2-explorando-bases

August 16, 2023

1 Tarea 2: Explorando bases

Francisco Mestizo Hernández A01731549

1.1 Instrucciones

1. Baja el archivo de trabajo
2. Analiza 2 de las siguientes variables en cuanto a sus datos atípicos y normalidad:

- Calorias
- Carbohidratos
- Proteinas
- Sodio
- Azucares (Sugars)

3. Para analizar normalidad se te sugiere:

- Realiza pruebas de normalidad univariada de las variables (selecciona entre los métodos vistos en clase)
- Grafica los datos y su respectivo QQPlot: qqnorm(datos) y qqline(datos) para cada variable
- Calcula el coeficiente de sesgo y el coeficiente de curtosis de cada variable.
- Compara las medidas de media, mediana y rango medio de cada variable.
- 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)
- Comenta los gráficos y los resultados obtenidos con vías a interpretar normalidad de los datos.

1.2 Inicialización del entorno

Iniciamos importando las librerias que necesitamos para usar algunas funciones como la que nos da la curtosis y el sesgo de error. Tambien importamos los datos desde el archivo csv que tenemos.

[78]:
```
#Instalamos las librerias que vamos a usar (para curtosis y coeficiente de␣
↪sesgo)
install.packages("moments")
```

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
```

[79]:
```
#Leemos los datos del csv (No los imprimo porque ocupan mucho espacio en la␣
 ↪pantalla)
M = read.csv('/content/sample_data/mc-donalds-menu-1.csv')
```

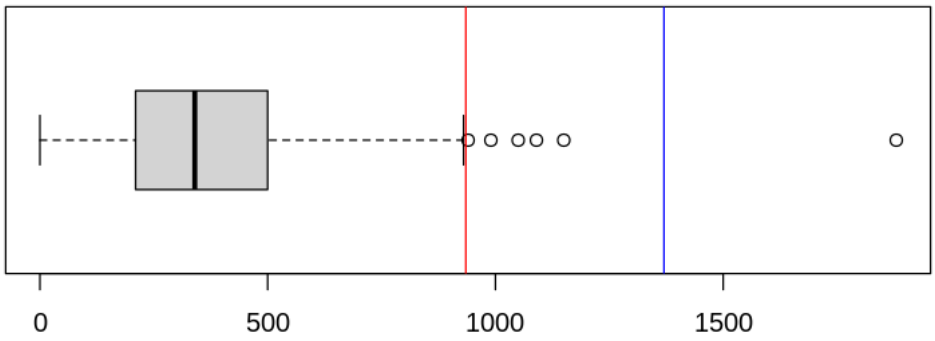
1.3 Evaluación de las calorias

Primero vamos a analizar la variable de calorías.

Con el código que se encuentra a continuación, hacemos el gráfico de caja y bigotes.

[80]:
```
#Evaluacion de los datos para las Calorias
X = M$Calories
q1=quantile(X,0.25)  #Cuantil 1 de la variable X
q3=quantile(X,0.75)  #Cuartil 3 de la variable X
ri= q3-q1             #Rango intercuartílico de X
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(X,horizontal=TRUE)
abline(v=q3+1.5*ri,col="red")  #linea vertical en el límite de los datos␣
 ↪atípicos
abline(v=q3+3*ri,col="blue") #linea vertical que marca os extremos
summary(X)
```

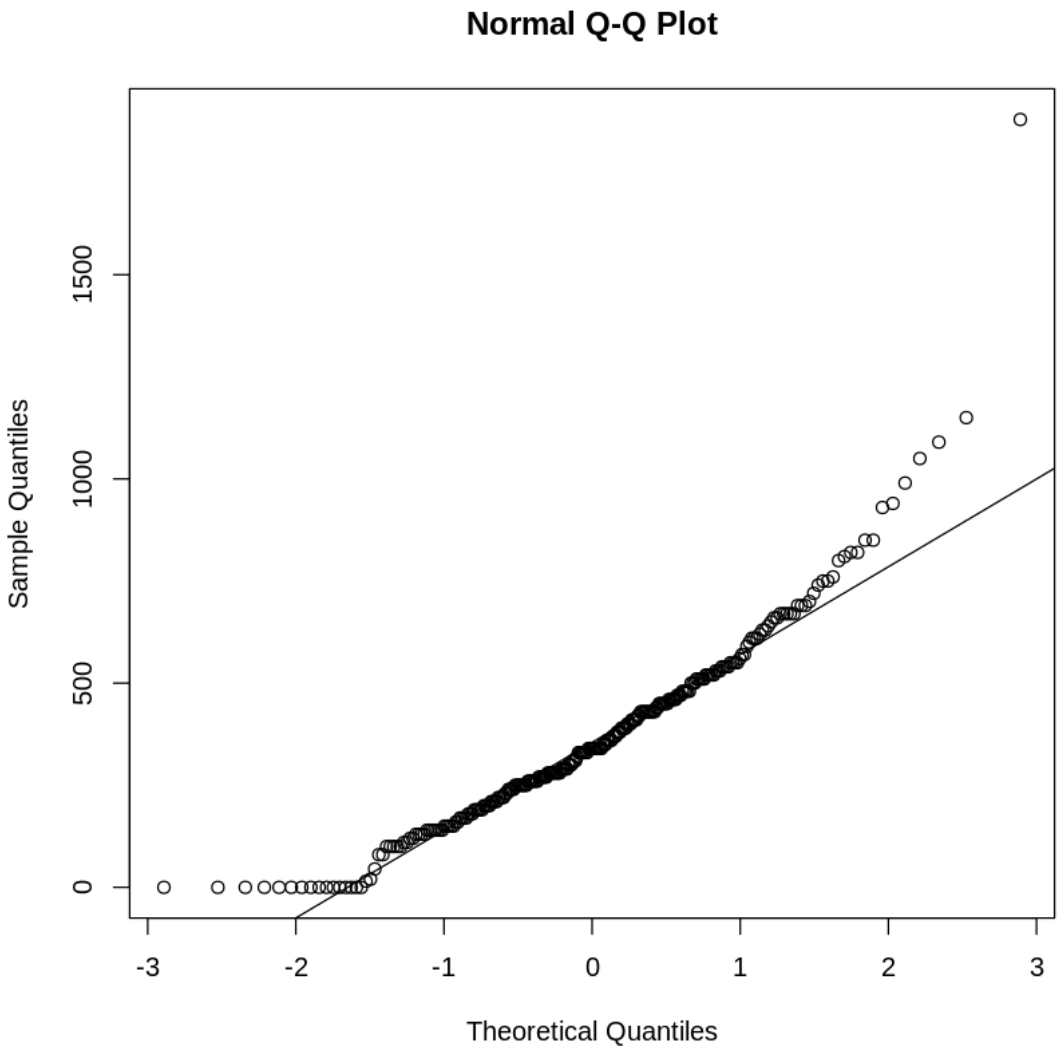
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   210.0   340.0   368.3   500.0  1880.0
```



Como resultado, después de la linea roja podemos ver los datos atípicos y después de la linea azul podemos ver los datos extremos. Esta variable si tiene algunos datos atípicos y solo un dato extremo.

Continuamos realizando la gráfica del QQPlot junto con su línea. Esta gráfica nos sirve para comprobar la normalidad de los datos.

[81]:
```
qqnorm(X) #Puntos normales para la qqplot
qqline(X) #Linea para la QQplot
```



En la QQPlot podemos ver que tenemos un sesgo positivo ya que los puntos de la cola tienden a irse para arriba de la linea que marcamos.

Ahora, podemos calcular el coeficiente de sesgo y la curtosis de la variable

[82]:
```
library(moments)
"Coeficiente de sesgo:"
moments::skewness(X) #Coeficiente de sesgo
"Curtosis: "
moments::kurtosis(X) #curtosis
```

'Coeficiente de sesgo:'

1.44410491051015

'Curtosis: '

8.64527387047867

Podemos ver que el sesgo es positivo y mucho mayor a uno, por lo que sabemos que los datos estan moderadamente sesgados a la derecha (se amontonan a la izquierda). Esto lo podemos confirmar con el histograma de más abajo.

Por otro lado, el resultado de la curtosis nos indica que la distribución está puntiaguda, ya que tiene un valor mayor a 3. Por lo tanto, podemos esperar una buena cantidad de datos atípicos.

Calculamos la media, mediana y rango de los datos

[83]:
```
"Media: "
mean(X)
"Mediana: "
median(X)
"Rango de los datos: "
range(X)
```

'Media: '

368.269230769231

'Mediana: '

340

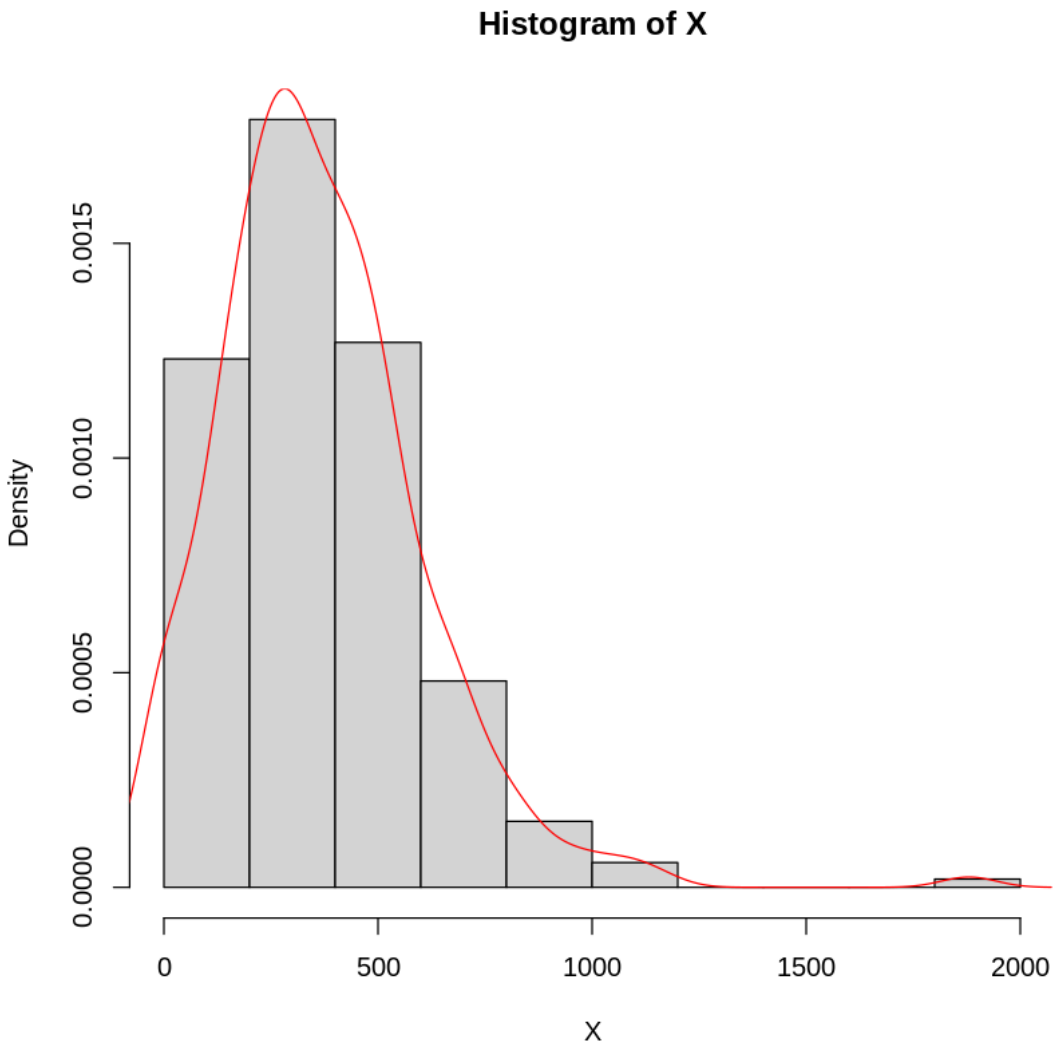
'Rango de los datos: '

1. 0 2. 1880

En los resultados de arriba vemos que los datos de las calorías van de 0 a 1880, el promedio de calorias por platillo es de 368 y la mediana es de 340.

Finalmente, graficamos el histograma de los datos, donde confirmamos que tiene un sesgo a la derecha, porque la media se encuentra un como más movida a la derecha que el punto más alto de la línea roja.

[84]:
```
hist(X,freq=FALSE)
lines(density(X),col="red")
```



1.4 Evaluación de las proteínas

Ahora, hacemos el mismo procedimiento pero con las proteínas.

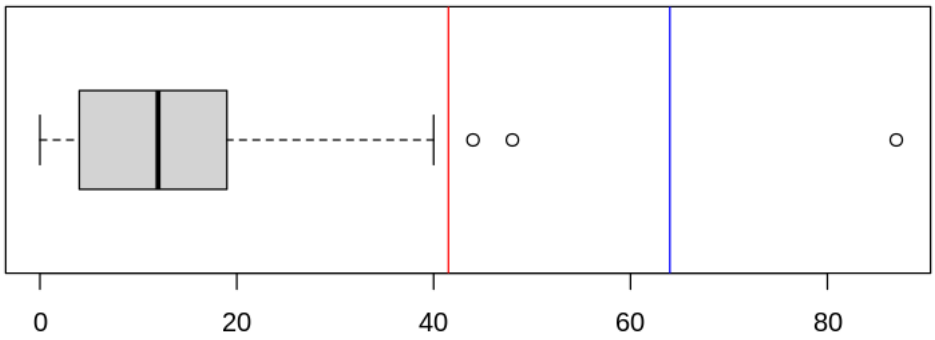
Comenazmos con el gráfico de caja y bigotes.

[85]:
```
#Evaluacion de los datos para las Proteinas

X = M$Protein
q1=quantile(X,0.25)  #Cuantil 1 de la variable X
q3=quantile(X,0.75)  #Cuartil 3 de la variable X
ri= q3-q1             #Rango intercuartílico de X
par(mfrow=c(2,1))  #Matriz de gráficos de 2x1
boxplot(X,horizontal=TRUE)
```

```r
abline(v=q3+1.5*ri,col="red")  #linea vertical en el límite de los datos␣
 ↪atípicos
abline(v=q3+3*ri,col="blue") #linea vertical que marca os extremos
summary(X)
```

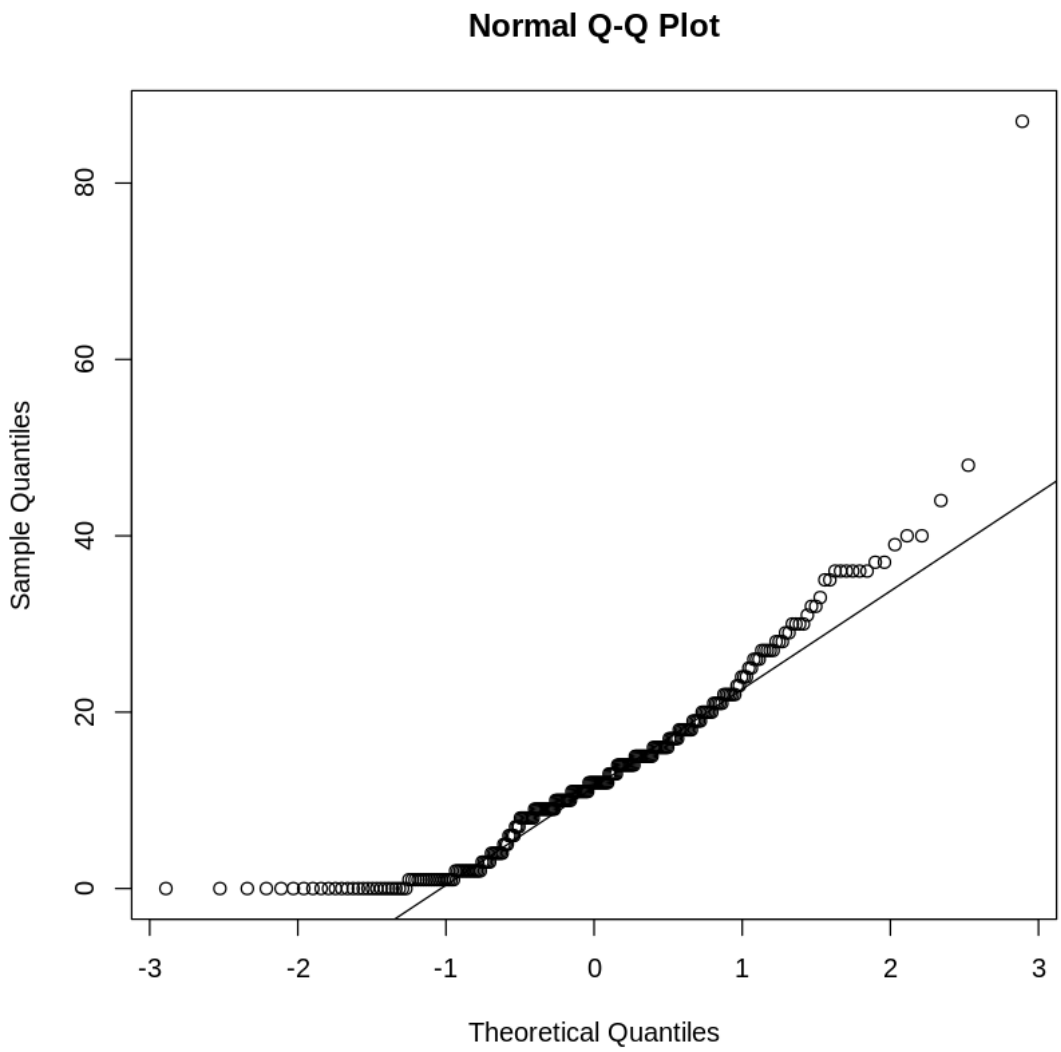
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    4.00   12.00   13.34   19.00   87.00
```



Podemos ver que en comparación con los datos de las calorías, tenemos muchos menos datos atípicos y también tenemos datos extremos.

Continuamos con el gráfico de la QQPlot.

[86]:
```
qqnorm(X) #Puntos de la QQplot
qqline(X) #Linea de la QQPlot
```



El comportamiento en la QQPlot podemos ver que es muy parecido a la anterior, por lo que existe un sesgo a la derecha en los datos. De todas formas, parece ser que las colas en esta gráfica son mucho más pronunciadas que en la QQPlot de las calorías.

Calculamos el coeficiente de sesgo y la curtosis:

[87]:
```
library(moments)
"Coeficiente de sesgo: "
moments::skewness(X)
"Curtosis: "
moments::kurtosis(X)
```

'Coeficiente de sesgo: '

1.57079418251428

'Curtosis: '

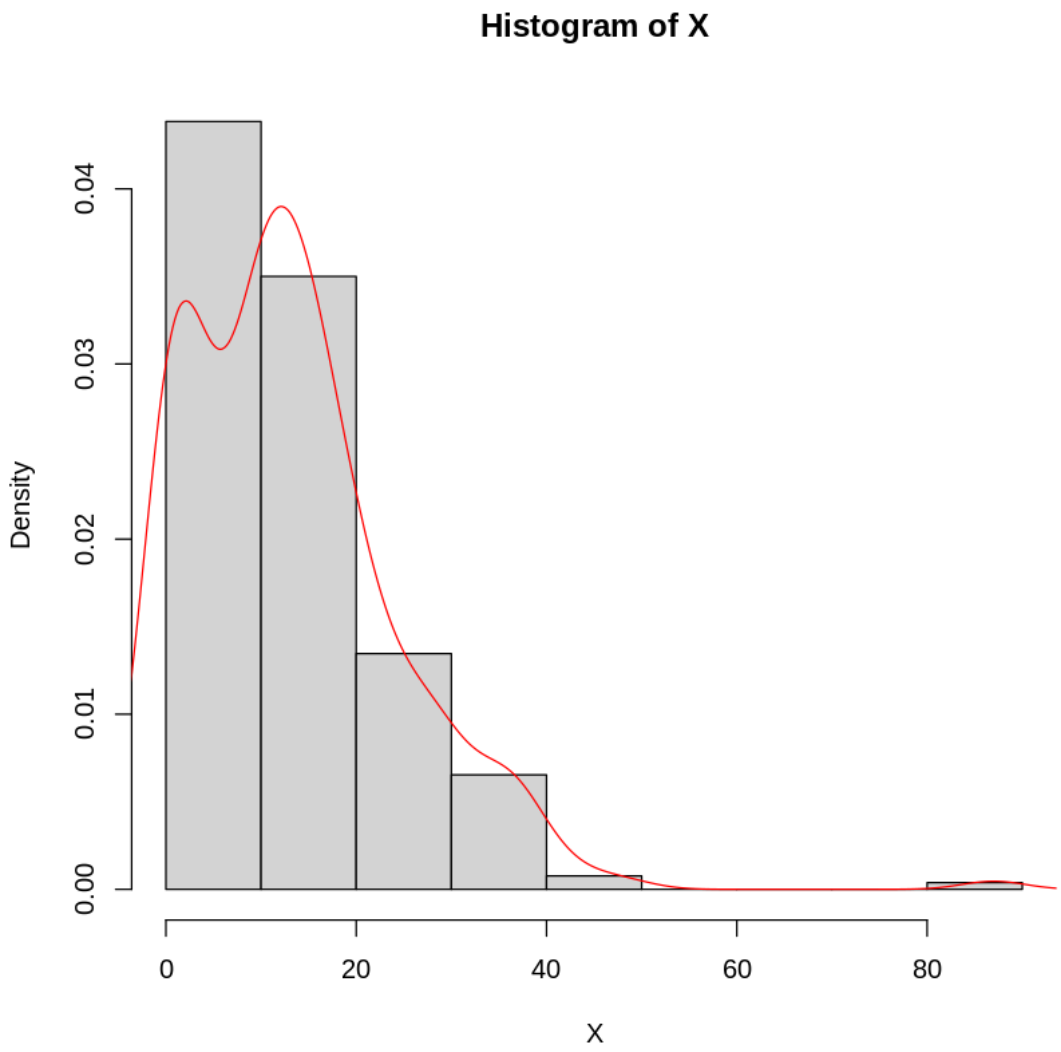
8.86354990872801

Como se puede observar que en la QQPlot tiene un comportamiento similar a las calorías, tambien este se presenta en el coeficiente de sesgo y la curtosis. Por los valores que obtuvimos, podemos decir que el sesgo es positivo y mucho mayor a uno, por lo que sabemos que los datos estan moderadamente sesgados a la derecha (se amontonan a la izquierda). Esto lo podemos confirmar con el histograma de más abajo.

Por otro lado, el resultado de la curtosis nos indica que la distribución está puntiaguda, ya que tiene un valor mayor a 3. Por lo tanto, podemos esperar una buena cantidad de datos atípicos.

Gráficamos el histograma

[88]:
```
hist(X,freq=FALSE)
lines(density(X),col="red")
```



En el histograma podemos ver el valor extremo y tambien el fuerte sesgo a la derecha que tiene la gráfica

1.5 Análisis y conclusiones

De acuerdo a los resultados obtenidos de estas variables, podemos ver que no tienen un comportamiento normal ya que se encuentran los dos muy sesgados a la derecha. Es muy probable que sean variables que estén muy relacionadas ya que tienen un mismo coeficiente de sesgo y de curtosis.

Comparando con otras variables del dataset, hay otras que cuentan con diferentes valores de curtosis y sesgo, pero aun así todas las gráficas presentaban un sesgo a la derecha.

En conclusión, los datos no son normales ya que su comportamiento no sigue el esperado de la campana de Gauss.

Para consultar el comportamiento en el notebook, consultar la siguiente liga: https://colab.research.google.com/drive/1ZDxIBS3u6M36Eya9YUeoDhfuZcOKUTD?usp=sharing