
Campus Monterrey

# Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Tarea 2. Explorando bases

Juan Pablo Castañeda Serrano A01752030

## *Reporte del Código*

El código proporcionado realiza un análisis exploratorio de un conjunto de datos, específicamente de una columna que contiene información sobre las calorías de un menú de McDonald 's. A continuación, se describe paso a paso lo que hace el código:

### *Importaciones:*

pandas: Biblioteca para manipulación y análisis de datos.
matplotlib.pyplot: Biblioteca para graficación.
numpy: Biblioteca para operaciones matemáticas avanzadas.
scipy.stats: Módulo para funciones estadísticas.

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import probplot, norm
from scipy.stats import skew, kurtosis
```

### *Lectura del Archivo:*

Se lee un archivo CSV llamado "mc-donalds-menu-1.csv" y se almacena en el DataFrame M.

```
M = pd.read_csv("mc-donalds-menu-1.csv")

X = M['Calories']
# 1. Cuartiles
q1 = X.quantile(0.25)
q3 = X.quantile(0.75)

y1 = X.min()
y2 = X.max()

# 2. Intercuartil
ri = q3 - q1

# 3. 2x1 grid
fig, axes = plt.subplots(2, 1)

# 4. boxplot
axes[0].boxplot(X, vert=False)
axes[0].set_xlim([y1, y2])

# 5. outliers
axes[0].axvline(x=q3 + 1.5 * ri, color='red')
```

### *Selección de Datos:*

Se selecciona la columna 'Calories' y se almacena en la variable X.

### Cálculo de Cuantiles:

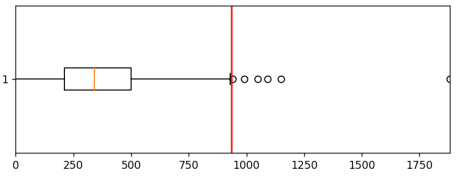
Se calcula el primer y tercer cuantil de X, almacenados en q1 y q3 respectivamente.
Se obtiene el mínimo (y1) y máximo (y2) valor de X.

### *Cálculo del Rango Intercuartílico (IQR):*

Se calcula el rango intercuartílico (IQR) restando q1 a q3 y se almacena en ri.

### *Visualización - Boxplot:*

Se establece una malla de gráficos de 2x1.
Se crea un boxplot de X en la primera fila.
Se dibuja una línea vertical en el límite para identificar datos extremos.




### *Filtrado de Datos Basados en Outliers:*

Se filtra el DataFrame M para retener sólo aquellos valores de 'Calories' que no sean considerados como outliers, basándose en el criterio de 1.5 * IQR por encima del tercer cuartil.

```
count    254.000000
mean     349.015748
std      201.401257
min        0.000000
25%      202.500000
50%      335.000000
75%      480.000000
max      930.000000
Name: Calories, dtype: float64
count    260.000000
mean     368.269231
std      240.269886
min        0.000000
25%      210.000000
50%      340.000000
75%      500.000000
max     1880.000000
Name: Calories, dtype: float64
```

***Resumen de Datos:***

Se imprime un resumen estadístico de los datos filtrados y los datos originales.

Q-Q Plot:

Se genera un Q-Q plot para visualizar si los datos siguen una distribución normal.

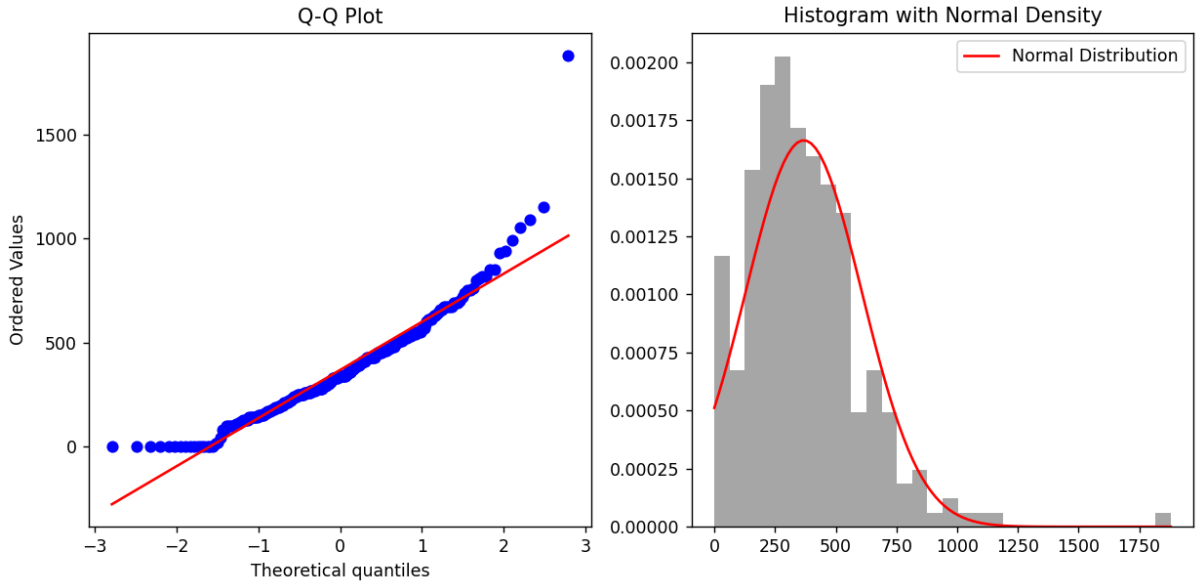
***Histograma con Curva de Densidad Normal:***

Se crea un histograma de X.

Se superpone una curva de densidad normal sobre el histograma.

***Skewness y Kurtosis:***

Se calculan y se imprimen la asimetría (skewness) y la kurtosis de X.




***Resultados / Conclusiones:***

El boxplot proporciona una vista rápida de la distribución de las calorías en el menú.

La línea roja en el boxplot ayuda a identificar potenciales valores atípicos.

El Q-Q plot y el histograma proporcionan información sobre la normalidad de los datos.

Los valores de skewness y kurtosis ofrecen medidas numéricas de la forma de la distribución.