DATA SCIENCE CLUB LAB 5
Population Data Analysis

Please take time to go through the dataset to effectively understand the questions.

Skills to gain:
- Data Preparation
- Basic Data Exploration
- Data Visualization
- Basic Statistics (Mean, Correlation, etc.)

Questions (Based on Kenya Population Dataset):

1. What is the mean, median, and standard deviation of the population densities across the counties?

2. Calculate the correlation between the total population and the number of households in each county. What does the correlation value suggest?

3. Create a bar plot to compare the total number of females and males for each county. What insights can you draw from this visualization?

4. Identify the county with the largest and smallest area (in square kilometers). How does their population density compare?

5. Plot a histogram of the population densities. Is the distribution skewed? If so, in which direction?

6. Calculate the ratio of males to females in each county. Which county has the highest male-to-female ratio?

7. Create a scatter plot to visualize the relationship between population density and the number of households. Do counties with higher population density tend to have more households?

8. Using box plots, identify any counties that are outliers in terms of total population. Which counties stand out, and why do you think that is?

9. Perform a simple linear regression between the area and total population of the counties. What can you conclude from the relationship?

10. Create a pie chart showing the proportion of the total population contributed by the top 5 most populous counties. How significant is the contribution of these counties compared to the others?