Project-1: Self-Kaggle dataset

Kaggle dataset:
https://www.kaggle.com/datasets/cpluzshrijayan/milkquality

About Dataset:

- This dataset consists of 7 independent variables pH, Temperature, Taste, Odor, Fat, Turbidity, and Color.
- Generally, the Grade or Quality of the milk depends on these parameters. These parameters play a vital role in the predictive analysis of the milk.
- The target variable is the Grade of the milk.
- It can be Low (Bad), Medium (Moderate), High (Good)
- If Taste, Odor, Fat, and Turbidity are satisfied with optimal conditions then they will assign 1 otherwise 0.
- The aim of the project is to determine the quality of milk based on the given parameters.
- The overall dataset contains 1059 rows and 8 columns.

Observations from the dataset:

- All the column names are proper except for Fat. Hence the extra space in the column name has been removed.
- The datatypes of all the columns are appropriate.
- There are no missing values in the dataset.

Univariate Analysis:

- pH:
    - The pH values range from 3.0 to 9.5, with an average of 6.63 and a standard deviation of 1.40.
    - The distribution slightly skews towards lower pH values based on a skewness of -0.68.
- Temperature:
    - The temperature values vary between 34.0 and 90.0, with a mean of 44.23 and a standard deviation of 10.10.

    - The temperature distribution is positively skewed (skewness of 2.22), indicating a longer tail towards higher temperatures.
- Taste:
    - Taste values are binary, 0 or 1, with a mean of 0.55 and a standard deviation of 0.50.
    - The data distribution slightly skews towards higher taste values (negative skewness of -0.19).
- Odor:
    - Odor values are also binary, 0 or 1, with a mean of 0.43 and a standard deviation of 0.50.
    - The distribution is slightly positively skewed (skewness of 0.27), indicating a tendency towards higher odor values.
- Fat:
    - Binary fat values (0 or 1) with a mean of 0.67 and a standard deviation of 0.47.
    - The data slightly skews towards lower fat values (negative skewness of -0.73).
- Turbidity:
    - Turbidity values range from 0.0 to 1.0, with a mean of 0.49 and a standard deviation of 0.50.
    - The distribution shows minimal skewness (0.04), suggesting a relatively balanced distribution.
- Colour:
    - Colour values vary from 240.0 to 255.0, with an average of 251.84 and a standard deviation of 4.31.
    - The distribution is negatively skewed (skewness of -1.02), indicating a tendency towards lower color values.
- Grade:
    - The 'Grade' attribute contains 1059 observations.
    - There are 3 unique categories: 'high', 'low', and 'medium'.
    - Value counts:
        - 'low': 429 occurrences.
        - 'medium': 374 occurrences.
        - 'high': 256 occurrences.

Bivariate Analysis:

Based on the provided information about the attributes of different grades of milk, the following observations can be made:

pH Levels:

- High and medium-grade milk have pH levels falling between 6 and 7.
- Low-grade milk encompasses a wider pH range, varying from 3 to 9.

Temperature:

- High and medium-grade milk have temperatures ranging between 35 and 50.
- Low-grade milk has a broader temperature range, from 30 to 70.

Taste:

- Low-grade milk is suggested to have a good taste compared to medium and high-grade milk.

Odor:

- High-grade milk is mentioned to have a good odor compared to low and medium-grade milk.

Fat Content:

- Low-grade milk contains a high amount of fat.
- Medium-grade milk contains both high and low fat in equal ratios.
- High-grade milk does not contain any fat

Turbidity:

- Low-grade milk is indicated to have high turbidity, while medium-grade milk has the least turbidity.

Color:

- Low and high-grade milk have color values within the range of 245 to 254.
- Medium-grade milk has a color range from 240 to 254.