**MSTA 1016: Statistics for Data Science**

**Assignment 1(20 Marks)**

1. Briefly describe the process of Data analysis from data gathering to data presentation. **(10 marks)**
2. You are tasked with the responsibility of assessing the extent of the effect of malaria in your community.
3. Briefly draft a problem statement in line with this assignment. **(2 marks)**
4. Create two objectives that will provide the solution to your problem statement. One objective must answer a qualitative question while another answers a quantitative question **(2 marks)**
5. For each objective, identify one dependent and 3 independent variables. **(4 marks)**
6. Categorize each variable and identify the units used to measure it. **(2 marks)**

**SOLUTION.**

**Number 1**

**Data Gathering**: This is the initial step in the data analysis process, where data is collected from various sources such as surveys, interviews, observations, or existing databases. The data is collected in a structured format, such as in a spreadsheet or database.

**Data Cleaning**: After data gathering, the data is cleaned to ensure that it is accurate, complete, and consistent. This involves checking for errors, handling missing values, and transforming data into a suitable format for analysis.

**Data Transformation**: In this step, the data is transformed into a format that is suitable for analysis. This may involve aggregating data, creating new variables, or converting data types.

**Data Analysis**: The transformed data is then analyzed using various statistical and data mining techniques to identify patterns, trends, and correlations. This may involve descriptive statistics, inferential statistics, data visualization, or machine learning algorithms.

**Data Interpretation**: The results of the data analysis are then interpreted to draw conclusions and make recommendations. This involves identifying the key findings, summarizing the results, and presenting the insights in a clear and concise manner.

**Data Presentation**: The final step is to present the results of the data analysis in a clear and concise manner, using visual aids such as charts, graphs, and tables. The presentation should be easy to understand and communicate the insights and findings to the target audience.

**Number 2.**

**a) Problem Statement**

Malaria is a public health challenge in our society, affecting individuals' health, productivity, and overall quality of life. This study aims to assess the extent and impact of malaria within the community by analyzing infection rates, patterns, and the socio-economic consequences of the disease. By gathering insights into malaria’s prevalence and its effects on various demographic groups, this assessment will inform local healthcare providers, community leaders, and public health officials to guide targeted interventions and resource allocation, ultimately contributing to malaria prevention and management efforts.

**b) Objectives**

1. **Qualitative Objective**: To understand the socio-economic impact of malaria on households in the community.
2. **Quantitative Objective**: To determine the prevalence rate of malaria within the community and identify patterns across different demographic groups.

**c) Variables**

1. **Qualitative Objective**
   * **Dependent Variable**: Socio-economic impact of malaria on households
   * **Independent Variables**:
     + Average household income (pre- and post-infection)
     + Number of days missed at work/school due to malaria
     + Access to healthcare facilities and treatment
2. **Quantitative Objective**
   * **Dependent Variable**: Malaria prevalence rate in the community
   * **Independent Variables**:
     + Age group
     + Gender
     + Proximity to stagnant water sources

**d) Categorization and Measurement Units**

| **Variable** | **Type** | **Category** | **Unit of Measurement** |
| --- | --- | --- | --- |
| Socio-economic impact | Dependent | Qualitative | Descriptive scale (e.g., low, medium, high) |
| Household income | Independent | Quantitative | Currency (USD, KES, etc.) |
| Days missed (work/school) | Independent | Quantitative | Days |
| Access to healthcare facilities | Independent | Qualitative | Binary (Yes/No) |
| Malaria prevalence rate | Dependent | Quantitative | Percentage (%) |
| Age group | Independent | Categorical | Age ranges (e.g., 0-5, 6-12, etc.) |
| Gender | Independent | Categorical | Male/Female |
| Proximity to stagnant water sources | Independent | Quantitative | Distance (meters) |