I. Problem Domain and Motivation (25 marks)

1. **Introduction to the Problem Domain**
    - **Definition**: Focus on the area of **Population and Society**, which includes demographics, social dynamics, and economic indicators affecting populations.
    - **Relevance**: Discuss how understanding population trends can inform policy decisions, social services, and community planning.
2. **Motivation**
    - **Choice of Dataset**: Explain the selection of a specific dataset (e.g., World Bank demographic data, census data, or survey data) that captures population characteristics over time.
    - **Implications**: Emphasize potential outcomes, such as identifying demographic shifts, understanding social inequalities, or informing public health initiatives.
3. **Dataset Chosen**
    - **Description**: Include the dataset's source, size, and attributes. For example:
        - **Source**: U.S. Census Bureau
        - **Size**: 10,000 rows and 10 columns
        - **Attributes**: Age, Gender, Income, Education Level, Employment Status, etc.
    - **Cleaning Confirmation**: Explain how you confirmed it meets the criteria (3000 rows, 7 columns) after initial cleaning.
4. **Challenges Faced**
    - **Missing Values**: Describe how many missing values were present and the strategies used to handle them (e.g., imputation, removal).
    - **Outliers**: Discuss methods for detecting outliers (e.g., Z-scores, IQR method) and your approach to managing them.
5. **Data Preparation Techniques**

Code Snippets:
r
Copy code

```r
# Load necessary libraries
library(dplyr)
library(ggplot2)

# Load the dataset
data <- read.csv("population_data.csv")

# Checking for missing values
missing_values <- colSums(is.na(data))
print(missing_values)
```

```
# Handling missing values
data <- na.omit(data)  # Simple removal
```

-