Title: Data Preprocessing: Cleaning and integrating datasets from multiple sources

1. Lab Objectives

- Understand and apply comprehensive data preprocessing techniques
- Handle real-world data quality issues using multiple methods
- Integrate datasets from multiple sources

2. Introduction

Data preprocessing is a critical step in any data science project. Real-world data often contains numerous quality issues that can significantly impact analysis results. This lab focuses on practical implementation of data preprocessing techniques and addressing common problems like missing values, inconsistencies, outliers, noise, duplicates, and data integration challenges.

The lab uses synthetic Nepal related survey dataset to demonstrate how proper preprocessing improves data quality and enables more reliable insights.

3. Datasets Description

Dataset Name	Description	Records
Household Survey Data	Synthetic data based on Nepal Multiple Indicator Cluster Survey with household demographics, income, education, and amenities	15 records
Education Facilities Data	District-wise education infrastructure including schools count, literacy rates, and student-teacher ratios	9 districts
Infrastructure Data	District-level development indicators including healthcare, road density, and electrification	9 districts

4. Preprocessing Tasks and Techniques

Task	Technique	Applied To	Example
Missing Value Handling	Mean/Median Imputation, Global Constant	monthly_income_npr, wealth_index	NaN → 107,714 (mean income)
Inconsistent Value Handling	Standardization, Mapping	district, education_level, water_source	"secondary" → "Secondary"
Outlier Handling	IQR Method, Winsorization	monthly_income_npr, family_size	1,200,000 → 116,750
Noisy Value Handling	Binning, Clustering	wealth_index, multiple variables	Wealth categories: Low/Medium/High
Duplicate Handling	Exact Matching	All columns	Removed household_id 14

5. Procedure

Step 1: Data Loading and Initial Assessment

- Load all three CSV files into pandas DataFrames
- Display basic information (shape, columns, data types)
- Identify missing values and data quality issues

Step 2: Missing Value Treatment

- Calculate mean for monthly_income_npr
- Calculate median for wealth_index
- Apply imputation to missing values
- Verify no missing values remain

Step 3: Data Standardization

- Standardize district names using str.title()
- Map education levels to consistent categories
- Standardize water source descriptions
- Convert boolean values to True/False

Step 4: Outlier Detection and Treatment

- Calculate IQR(or interquartile range, is a measure of statistical dispersion that shows the spread of the middle 50% of a dataset.) for monthly_income_npr and family_size
- Identify outliers beyond 1.5×IQR range
- Apply winsorization to cap extreme values
- Visualize before/after distributions

Step 5: Noise Reduction

- Create wealth categories using binning (Low/Medium/High/Very High)
- Apply K-means clustering for data validation
- Assign cluster labels to each household

Step 6: Duplicate Removal

- Identify exact duplicate records
- Remove duplicated household entries
- Verify unique household_ids

Step 7: Data Integration

- Merge household data with education facilities data
- Merge result with infrastructure data
- Verify integrated dataset structure

6. Results and Comparison

Metric	Before Preprocessing	After Preprocessing
Dataset Size	15 records	14 records (1 duplicate removed)
Missing Values	3 missing values	0 missing values
Average Income (NPR)	107,714 (with outlier)	58,452 (outlier treated)
Data Consistency	Mixed cases, inconsistent categories	Standardized values

7. Discussion Questions (to be submitted handwritten)

Q1. What are the main datasets used in this lab. List the data points which requires cure in these dataset. (Hint: `household_id 2`: `monthly_income_npr` = `NaN`)

Q2. Why is data preprocessing important before any analysis? List what preprocessing techniques you applied for each of the problems listed in Q1.

Q3. How did outlier treatment affect the average income calculation? Which value is more representative?

Q4. What are the advantages and disadvantages of using mean vs median for missing value imputation?

Q5. How does data integration from multiple sources enhance analysis capabilities?

Q6. What additional preprocessing steps might be needed for real-world survey data?

8. Submission Guidelines

- Complete the Jupyter notebook with all preprocessing steps implemented. Document each step with comments and explanations. Submit the well documented notebook file (.ipynb) to bidur(@)gces.edu.np with email subject: BECE2022 - CMP 360 – Lab#1. The email body must have your: Name, Class Roll Number, Lab Number and Lab Title. Use your gces email to complete the submission.
- Hardcopy Submission (Individual **Handwritten**):
  - Lab Title
  - Objectives
  - Discussion Questions & Answers