Data Preprocessing

Cleaning and Integrating Datasets from Multiple Sources

Course: AI and Tourism – MIT-AI @ Gandaki University

Bidur Devkota, PhD
GCES Pokhara

January 20, 2026

Introduction

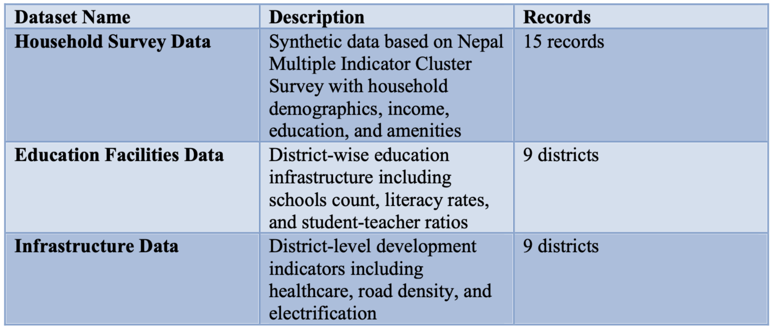
Data preprocessing is a critical step in any machine learning and data science projects.

- Real-world data often contains quality issues
- Impacts analysis results significantly
- This lab focuses on practical implementation of preprocessing techniques
- Uses synthetic survey dataset to demonstrate improvements

Objectives

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

Dataset Overview

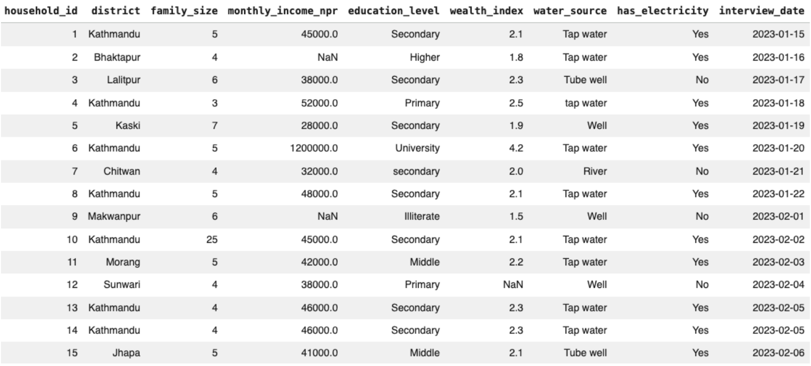


Dataset Name	Description	Records
Household Survey Data	Synthetic data based on Nepal Multiple Indicator Cluster Survey with household demographics, income, education, and amenities	15 records
Education Facilities Data	District-wise education infrastructure including schools count, literacy rates, and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

Figure: Sample survey dataset

Data : Click to Download

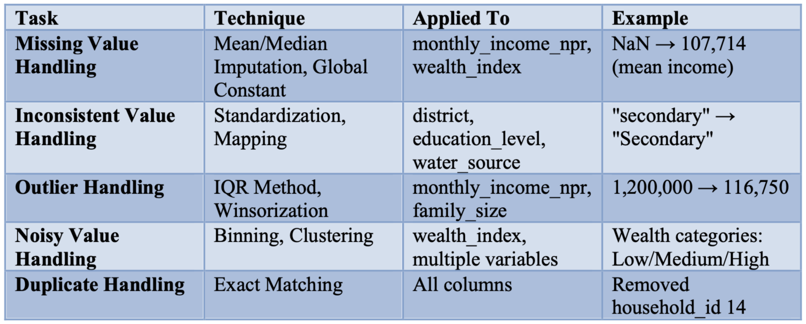
Dataset Overview



household_id	district	family_size	monthly_income_npr	education_level	wealth_index	water_source	has_electricity	interview_date
1	Kathmandu	5	45000.0	Secondary	2.1	Tap water	Yes	2023-01-15
2	Bhaktapur	4	NaN	Higher	1.8	Tap water	Yes	2023-01-16
3	Lalitpur	6	38000.0	Secondary	2.3	Tube well	No	2023-01-17
4	Kathmandu	3	52000.0	Primary	2.5	tap water	Yes	2023-01-18
5	Kaski	7	28000.0	Secondary	1.9	Well	Yes	2023-01-19
6	Kathmandu	5	1200000.0	University	4.2	Tap water	Yes	2023-01-20
7	Chitwan	4	32000.0	secondary	2.0	River	No	2023-01-21
8	Kathmandu	5	48000.0	Secondary	2.1	Tap water	Yes	2023-01-22
9	Makwanpur	6	NaN	Illiterate	1.5	Well	No	2023-02-01
10	Kathmandu	25	45000.0	Secondary	2.1	Tap water	Yes	2023-02-02
11	Morang	5	42000.0	Middle	2.2	Tap water	Yes	2023-02-03
12	Sunwari	4	38000.0	Primary	NaN	Well	No	2023-02-04
13	Kathmandu	4	46000.0	Secondary	2.3	Tap water	Yes	2023-02-05
14	Kathmandu	4	46000.0	Secondary	2.3	Tap water	Yes	2023-02-05
15	Jhapa	5	41000.0	Middle	2.1	Tube well	Yes	2023-02-06

Figure: Sample Household Survery Data

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

Figure: Preprocessing pipeline overview

Methodology: Step 1 – Data Loading and Initial Assessment

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

Methodology: 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

Methodology: Step 3 – Data Standardization

Field	Standardization Rule
District names	`str.title()`
Education levels	Mapped to consistent categories
Water source	Standardized descriptions
Boolean values	Converted to True/False

Table: Data standardization steps applied

Methodology: Step 4 – Outlier Detection and Treatment

Column	IQR	Outlier Threshold (1.5×IQR)
monthly_income_npr	[IQR value]	[Threshold]
family_size	[IQR value]	[Threshold]

Table: IQR and outlier thresholds

- The **Interquartile Range (IQR)** measures *data spread* by showing the range of the middle 50% of values, calculated as the difference between the third quartile and the first quartile *(IQR=Q3Q1)*.
- Usefulness: IQR is **resistant to extreme outliers**, making it a reliable indicator of where the majority of the data is concentrated.

Methodology: Step 5 – Noise Reduction

- Created wealth categories using binning (Low/Medium/High/Very High)
- Applied K-means clustering for data validation
- Assigned cluster labels to each household

Methodology: Step 6 – Duplicate Removal

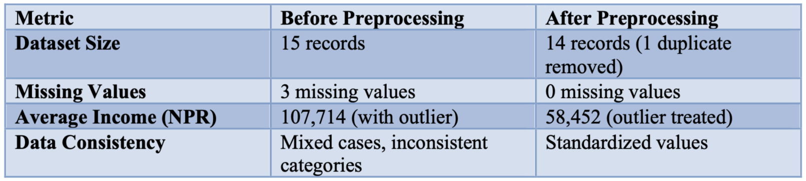
Action	Count
Exact duplicate records identified	[Number]
Duplicated household entries removed	[Number]
Unique household_ids after removal	[Number]

Table: Duplicate removal summary

Methodology: Step 7 – Data Integration

- Merged household data with education facilities data
- Merged result with infrastructure data
- Verified integrated dataset structure

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

Figure: Preprocessing results comparison

Discussion Questions

1. What are the main datasets used in this lab? List the data points which require cure.
2. Why is data preprocessing important before any analysis? List techniques applied.
3. How did outlier treatment affect the average income calculation?
4. Advantages and disadvantages of using mean vs median for imputation?
5. How does data integration enhance analysis capabilities?
6. What additional preprocessing steps might be needed for real-world survey data?

Submission Guidelines

- Complete Jupyter notebook with all preprocessing steps
- Document each step with comments and explanations
- Submit .ipynb file via email
- Subject: TAI2025 – Tourism and AI – Lab#4
- Email body: Name, Class Roll Number, Lab Title

Required Submission Structure

- Lab Title
- Objectives
- Methodology
- Discussion Questions & Answers
- Conclusion

Thank You

Questions?