1. Introduction

The objective of this analysis is to examine household consumption expenditure patterns using survey data collected by the National Sample Survey Office (NSSO) in India. The analysis aims to achieve two key tasks: (1) calculating the national share of spending across different product categories classified under COICOP Level 1, and (2) computing and visualizing income inequality using the Lorenz curve and Gini coefficient. The insights derived from this study help in understanding expenditure distribution across different household demographics and evaluating economic disparity.

2. Dataset

The analysis utilizes three primary datasets:

- households.csv: This file contains household-level information, including a unique household identifier (hh_id), survey weight (weight), whether the household is in an urban or rural region (urban), and the number of members in the household (hh_size).
- expenses.csv: This file records the annual expenditure of each household (hh_id) on various products (product_id).
- products.csv: This file maps each product (product_id) to its corresponding category in the COICOP 1999 classification, structured into four hierarchical levels (coicop_survey_1, coicop_survey_2, coicop_survey_3, coicop_survey_4).

These datasets, located in the input_data directory, provide the necessary information to evaluate household spending behavior and income distribution.

3. Methodology

The data processing is divided into two main tasks:

Task 1: National Share of Spending Calculation

1. Load the datasets and merge them appropriately using hh_id and product_id as keys.
2. Multiply household expenditures by survey weights to obtain a weighted expenditure value.
3. Map the categories code to COICOP group names
4. Aggregate the weighted expenditures across different COICOP Level 1 categories.

5. Normalize the values by the total expenditure to compute the percentage share of spending in each category.

Task 2: Lorenz Curve and Gini Coefficient Calculation

1. Compute the total household expenditure by aggregating expenditures across products.
2. Calculate per capita household expenditure by dividing total expenditure by household size (hh_size).
3. Sort households by per capita expenditure in ascending order.
4. Compute cumulative population share and cumulative expenditure share, incorporating household weights.
5. Plot the Lorenz curve using these cumulative shares.
6. Compute the Gini coefficient as the ratio of the area between the Lorenz curve and the 45-degree equality line to the total area under the equality line.

Definitions:

- COICOP (Classification of Individual Consumption by Purpose): A hierarchical system used to classify consumption expenditures into categories.
- Lorenz Curve: A graphical representation of the cumulative distribution of income or expenditure.
- Gini Coefficient: A statistical measure of income inequality, ranging from 0 (perfect equality) to 1 (maximum inequality).

4. Resources

- Code Repository: code
- Dataset Location: Dataset
- Notes: The analysis relies on pandas, NumPy, and Matplotlib for data manipulation and visualization. Ensure that the dataset files are correctly placed in the designated directory before running the script.