Group 8 - Data Analysis and Visualization Preliminary Progress Report

Abrigo, Nathanael Chris O. Daliuag, Ronan Manuel
Padrejuan, Shaun Kristoffer C. Salvador, Fabian III R.
4CSC

I. Introduction

The Philippine Statistics Authority (PSA) is the primary agency responsible for collecting and providing statistics data for the Philippines. One of its surveys is the Family Income and Expenditure Survey (FIES), which offers data involving the income sources, spending, and living conditions of Filipino households.

Our project utilizes this dataset to analyze family income and expenditure patterns, focusing on key factors such as regional disparities, household consumption, housing expenses, and more. By examining this data, we aim to provide deeper insights into economic behaviors that can inform decision-making for both policymakers and investors with the use of visualization techniques such as bar graphs and the like.

II. Problem Statement and Objectives

Analyzing the FIES dataset involves the analysis of the total household incomes for every source of income and the total household expenditures for every source of expenditure. The objectives are to find patterns and trends that stand out in terms of the relationship between income and expenditure, the relationship and contribution of income sources and expenditure sources to the total values of both, and to formulate Machine Learning Models to use the FIES dataset to either predict or classify a variable like income groups and to cluster the dataset to visualize the household demographic of the 2023 FIES dataset.

For the Machine Learning Model, two use cases were proposed. A classification artificial neural network to classify households based on the seven income groups from the feature engineering using expenditures to see if high expenditures mean higher income, the most rational way of classification, and a DBSCAN clustering to have a much more complex clustering algorithm and to also account for similar households in terms of income but not for expenditures.

III. Methodology

The first phase includes data cleaning and data validation. Null values were found on the Total Household Disimbursements column as per the data dictionary, which has an object datatype on the column. While the derivations of the column were found by hand, the group has presently decided to drop the null value columns to preserve data integrity.

An unnamed_13 column was also found, which had no column name but is part of the derivations of the other columns, so the group decided to add its value to the other income_nec_column and drop the column itself to have standardized data.

Feature Engineering was conducted to have a classification variable for machine learning and income and expenditure features. Other variables include Economic Stability and Consumption Patterns. Summary statistics were also used for the numerical columns, with the exception of the ID columns.

Data Visualization includes exploring the Mean and Total Expenditure per category using a bar plot to gauge the expenditure levels of all categories. The house rent per region was also graphed in a bar graph to show the difference in rental expenditure per region in the Philippines. Mean Total Income vs. Total Disbursements by Family Size was graphed with a bar graph as well to show the difference between different family sizes as per the metadata of the PSA. The average monthly income per region was also graphed in a bar graph. Several visualizations were made on the disbursement of alcohol and tobacco, including family size, per capita income decile, and urban/rural areas. The average agricultural income by region uses a multi-variate bar graph to check all the income sources and what works per region. The average difference between the total household income and expenditures by per capita income decile shows the potential savings of households divided by deciles and their differences. A pie chart visualized the mean expenditure of food consumed at home vs. outside to get a view of the two variables and how households plan their food expenditure. An expenditure breakdown of income from salaries/wages and no income from the former was made with a bar graph, showcasing the things families with stable income but that families with no stable income cannot buy. A pie chart showcasing the food vs non-food expenditures overall shows whether or not non-food expenditures eclipse the spending on the more essential food expenditures. A pie chart representing education expenditure in Rural and Urban Areas was also made. A heat map was also utilized to correlate non-food expenditures.