HOUSEHOLD INCOME CLASS PREDICTION IN THE PHILIPPINES THROUGH HOUSEHOLD EXPENDITURES

By

CHRISTIAN DAVE O. EMPINADO

PETE ANDRE C. CADELINA

Submitted to

GLENN PEPITO

January 5, 2023

INTRODUCTION

Income is a determined factor on the quality of life a household, family, or individual may have. With the economic landscape of the Philippines today, determining one's potential household income can provide insight on policy making and economic ventures in the Philippines. According to a study by Mark Wooden (2004), income can help determine the " economic circumstances of a household." By examining what households spend their money on and the assets they own, it is possible to make predictions about their income level.

There are several benefits in predicting household income. It can help understand patterns of inequality within a population. Through data, factors on inequality may be determined and policies may be enforced. Another benefit is with planning and budgeting. The government can prepare accordingly for different income groups. A higher budget may be needed to support low-income groups and vice versa. Businesses can also benefit from such predictions. They may use income data to target marketing efforts and tailor their products and services to income groups in specific locations.

It is possible to predict household income through machine learning models. The study of Matkowski (2021) showed that a machine learning approach provides better performance in predicting income compared to traditional prediction approaches. The study also noted that there has been an increase in false reporting of income. Individuals and households have been declaring their income lower than their actual income.

There have been occurrences in the Philippines of false reporting of income as well. The study of Villejo (2014) stated that some people declare a lower income for the benefits. This leads to costly and time consuming accurate assessment. Hence, the study of Benin & Randriamamonjy (2008) uses expenditure to estimate household income changes in a set period of time.

Studies have shown that income does not determine the economic welfare of a household. It is discovered that consumption or expenditures are a better measure for economic welfare (Berhanu, 1999). Hence, they may be used as factors to determine income needed for an appropriate economic welfare.

Other studies have used expenditures to predict household income. This prediction may be used by the government for policies and better management. The study of Sri (2021). Used decision tree and random forest regression algorithms to determine household income using family expenses. This is useful in gaining insight on the needed income for a certain expenditure amount. Another study used electricity consumption as a predictor of household income. Electricity consumption was used to classify and characterize families. It is shown that electricity consumption may be an efficient predictor of income (Francisco et al., 1970).

Predicting household income through expenditures can provide valuable insights as to how much income is needed for certain lifestyles and expenditures. It can help identify areas that are in need of services and advise on the planning and budgeting of resources. This research aims to explore the use of data of household expenditures and possessions in predicting household income. May this research help aid in the reduction of income inequality and provide a deeper understanding on the economic landscape of the Philippines.

BUSINESS UNDERSTANDING

Poverty is a persistent problem in the Philippines, and the poverty threshold is one metric in determining if a household is poor. This is used to determine the poverty rate of the country. A household is considered poor if its level of income or consumption is below the poverty threshold. The government determines the poverty threshold through the cost of basic needs and services and through Family Income and Expenditure Survey, which is conducted every three years by the Philippine Statistics Authority.

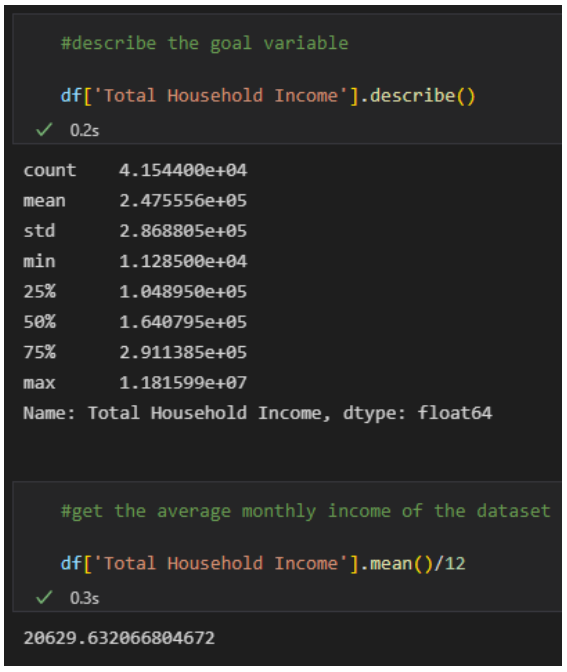
The Family Income and Expenditure Survey contains data on several expenditures a household may have such as housing and water expenditures, communication expenditure, and total food expenditure. It also contains data on possessions and house material quality. Aside from these, data on the main source of income, household head sex, household head occupation, total number of family members, etc.

With this data, it is possible to make predictions on household income needed for expenditures. This can be done through expenditures being a factor of household income. Insights may be provided on the disparity of income in a population, and patterns may be formed. Further research on concerning topics like poverty and income inequality may be conducted as well.

DATA DESCRIPTION

The dataset contains 60 columns with 41,544 rows. An ID column was added to establish differences for visualization through Power BI. There are 46 numerical columns including the ID column and 15 categorical columns. Exploratory Data Analysis was performed to further study the dataset. The chosen goal variable is 'Total Household Income' which will be used to predict the income class.

The goal variable was first examined. The Total Household Income represents the yearly income of the household. The average household income from the dataset is PHP 247,555.60 divided by 12 months, that is PHP 20,629.63 per month. The median income is at PHP 164,079.50 per year, or PHP13,673.29 per month. The highest income is PHP 11,815,990.00 per year, or PHP 984,665.00 per month. The lowest income however is just PHP 11,285.00 per year, or PHP 940 per month.



```
#describe the goal variable

df['Total Household Income'].describe()
```

✓ 0.2s

```
count    4.154400e+04
mean     2.475556e+05
std      2.868805e+05
min      1.128500e+04
25%      1.048950e+05
50%      1.640795e+05
75%      2.911385e+05
max      1.181599e+07
Name: Total Household Income, dtype: float64
```

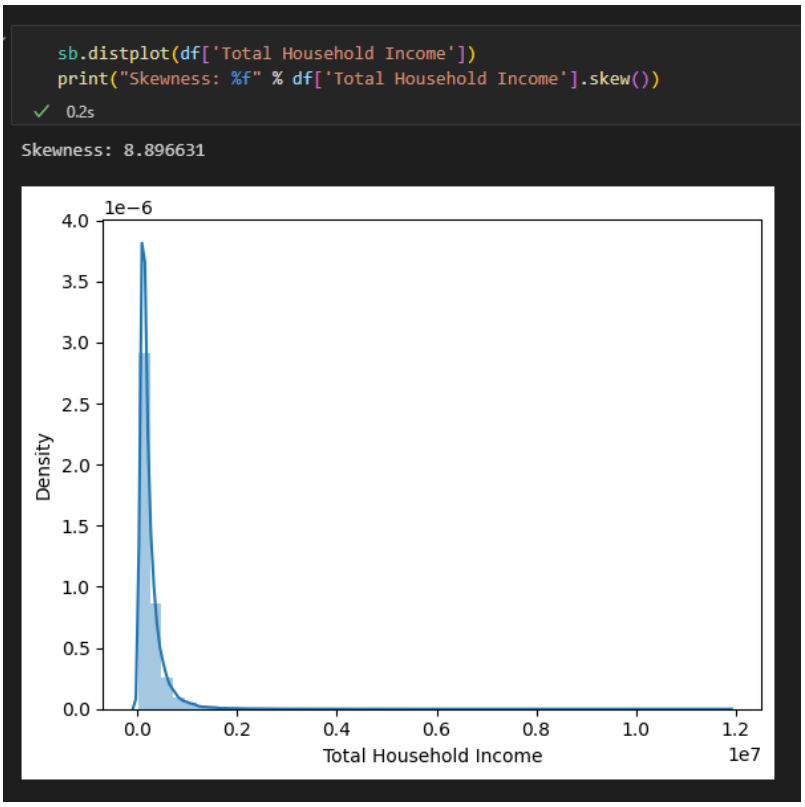
```
#get the average monthly income of the dataset

df['Total Household Income'].mean()/12
```

✓ 0.3s

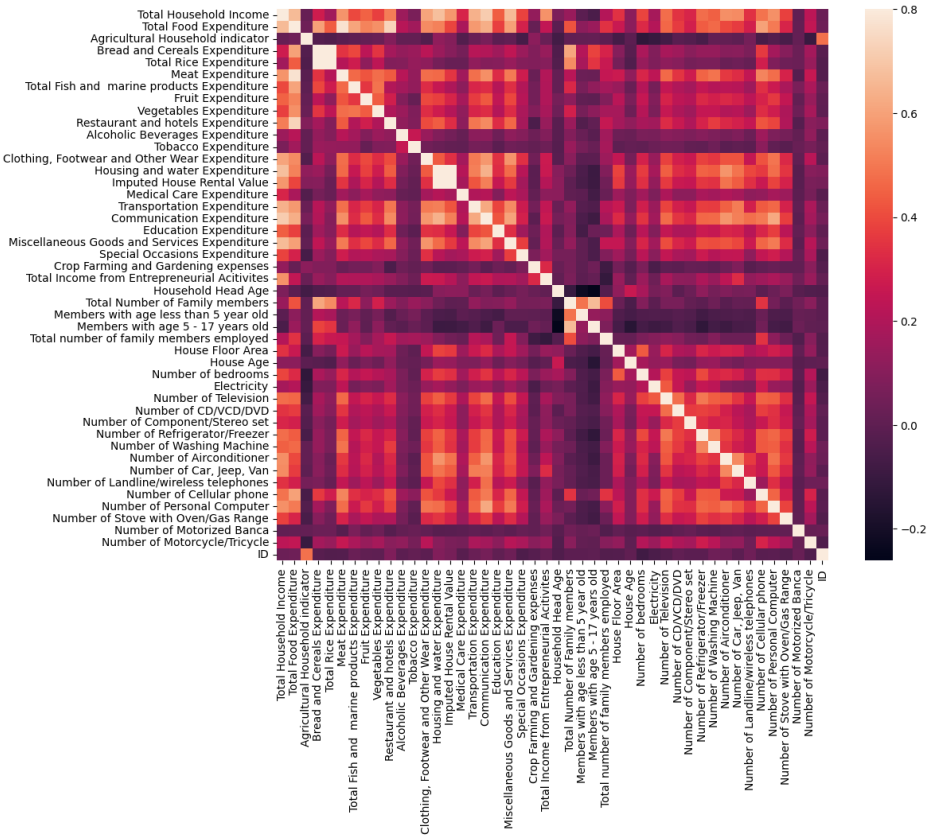
```
20629.632066804672
```

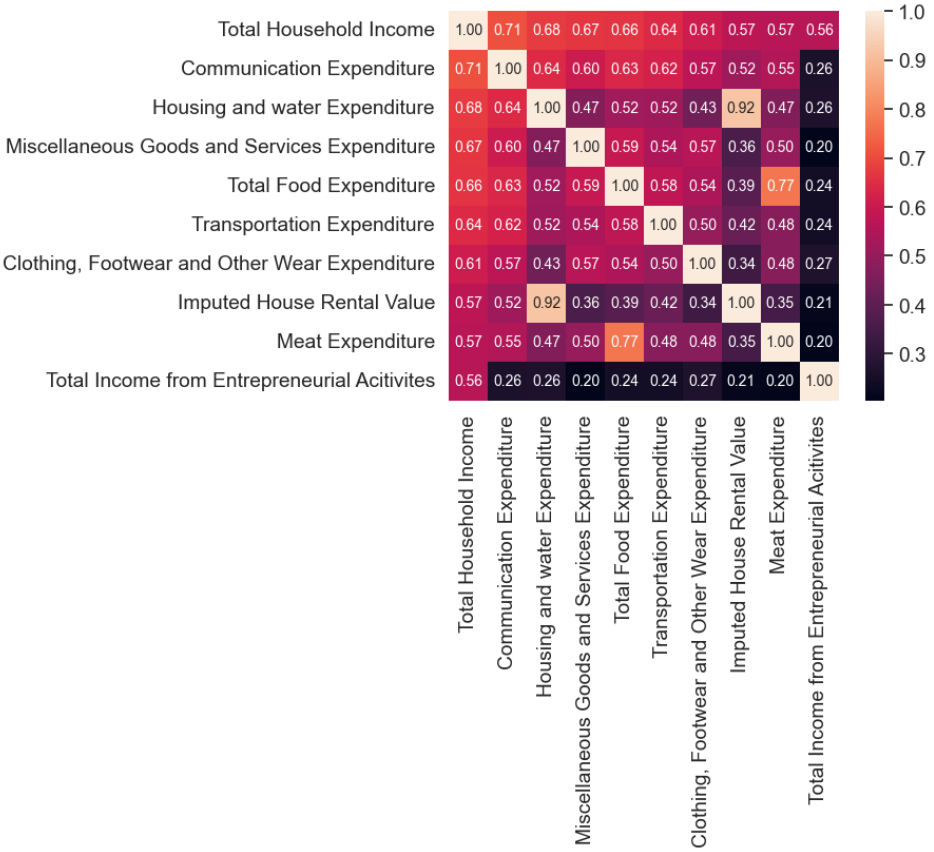
A distribution graph was created. The data clearly deviates from a normal distribution, and shows that the data is heavily skewed to the left and is positively skewed.

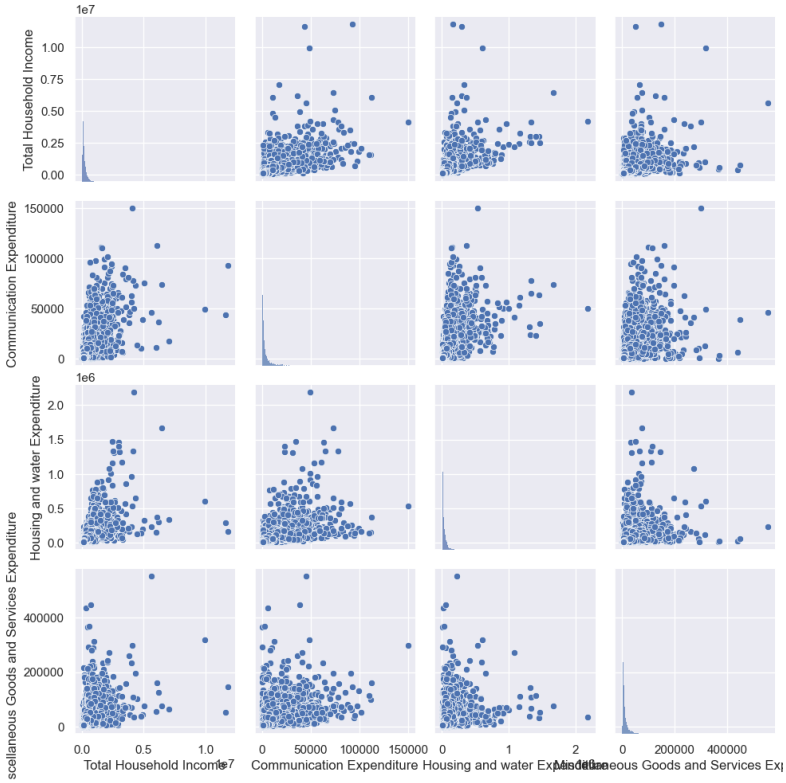


A correlation matrix was then established to see the goal variable's relationship with other fields. The 3 variables with the highest correlation were "Communication Expenditure" with 0.71, "Housing and water Expenditure" with 0.68, and "Miscellaneous Goods and Services Expenditure" with 0.67.

Scatter plots show that outliers are present within the dataset. It is also observed that there are no variables with clear positive correlation between the goal variable; however, clusters have been observed denoting that lower expenditure is noticed on all income classes. This also means that the variables show a huge variance which will limit the reliability of a regression model.

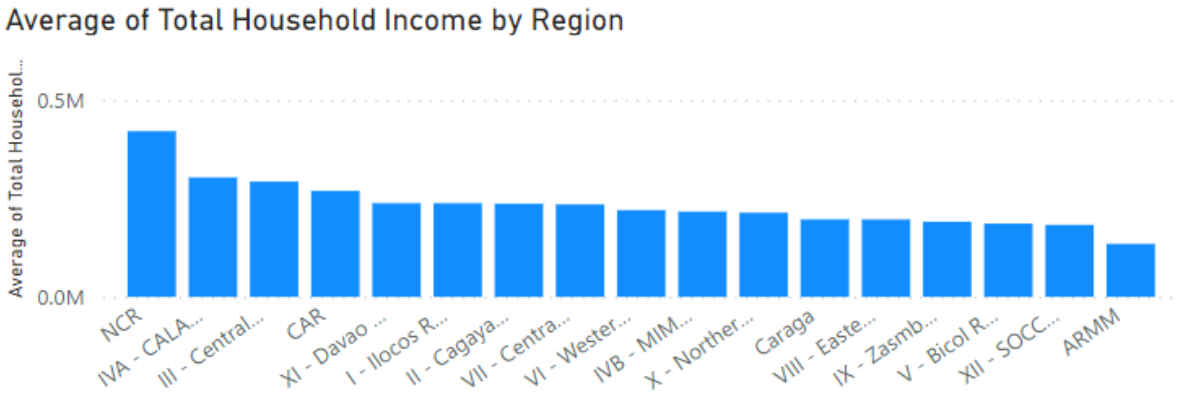

Total Household Income
Total Food Expenditure
Agricultural Household indicator
Bread and Cereals Expenditure
Total Rice Expenditure
Meat Expenditure
Total Fish and marine products Expenditure
Fruit Expenditure
Vegetables Expenditure
Restaurant and hotels Expenditure
Alcoholic Beverages Expenditure
Tobacco Expenditure
Clothing, Footwear and Other Wear Expenditure
Housing and water Expenditure
Imputed House Rental Value
Medical Care Expenditure
Transportation Expenditure
Communication Expenditure
Education Expenditure
Miscellaneous Goods and Services Expenditure
Special Occasions Expenditure
Crop Farming and Gardening expenses
Total Income from Entrepreneurial Acitivites
Household Head Age
Total Number of Family members
Members with age less than 5 year old
Members with age 5 - 17 years old
Total number of family members employed
House Floor Area
House Age
Number of bedrooms
Electricity
Number of Television
Number of CD/VCD/DVD
Number of Component/Stereo set
Number of Refrigerator/Freezer
Number of Washing Machine
Number of Airconditioner
Number of Car, Jeep, Van
Number of Landline/wireless telephones
Number of Cellular phone
Number of Personal Computer
Number of Stove with Oven/Gas Range
Number of Motorized Banca
Number of Motorcycle/Tricycle
ID
0.8
0.6
0.4
0.2
0.0
−0.2


Total Household Income 1.00 0.71 0.68 0.67 0.66 0.64 0.61 0.57 0.57 0.56
Communication Expenditure 0.71 1.00 0.64 0.60 0.63 0.62 0.57 0.52 0.55 0.26
Housing and water Expenditure 0.68 0.64 1.00 0.47 0.52 0.52 0.43 0.92 0.47 0.26
Miscellaneous Goods and Services Expenditure 0.67 0.60 0.47 1.00 0.59 0.54 0.57 0.36 0.50 0.20
Total Food Expenditure 0.66 0.63 0.52 0.59 1.00 0.58 0.54 0.39 0.77 0.24
Transportation Expenditure 0.64 0.62 0.52 0.54 0.58 1.00 0.50 0.42 0.48 0.24
Clothing, Footwear and Other Wear Expenditure 0.61 0.57 0.43 0.57 0.54 0.50 1.00 0.34 0.48 0.27
Imputed House Rental Value 0.57 0.52 0.92 0.36 0.39 0.42 0.34 1.00 0.35 0.21
Meat Expenditure 0.57 0.55 0.47 0.50 0.77 0.48 0.48 0.35 1.00 0.20
Total Income from Entrepreneurial Acitivites 0.56 0.26 0.26 0.20 0.24 0.24 0.27 0.21 0.20 1.00
Total Household Income
Communication Expenditure
Housing and water Expenditure
Miscellaneous Goods and Services Expenditure
Total Food Expenditure
Transportation Expenditure
Clothing, Footwear and Other Wear Expenditure
Imputed House Rental Value
Meat Expenditure
Total Income from Entrepreneurial Acitivites
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3

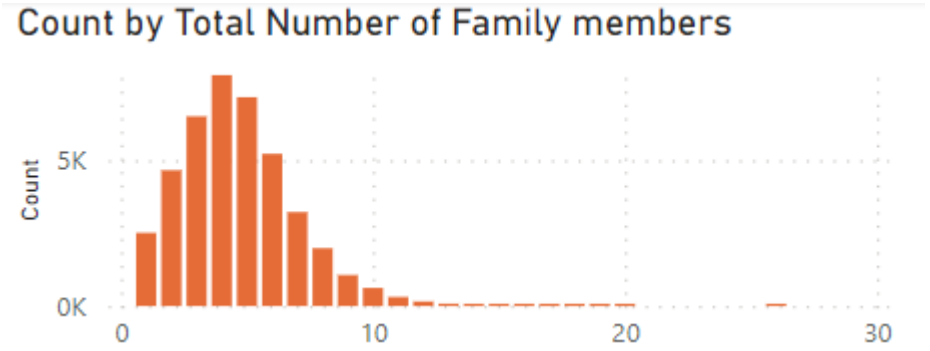


ANALYSIS AND RESULTS

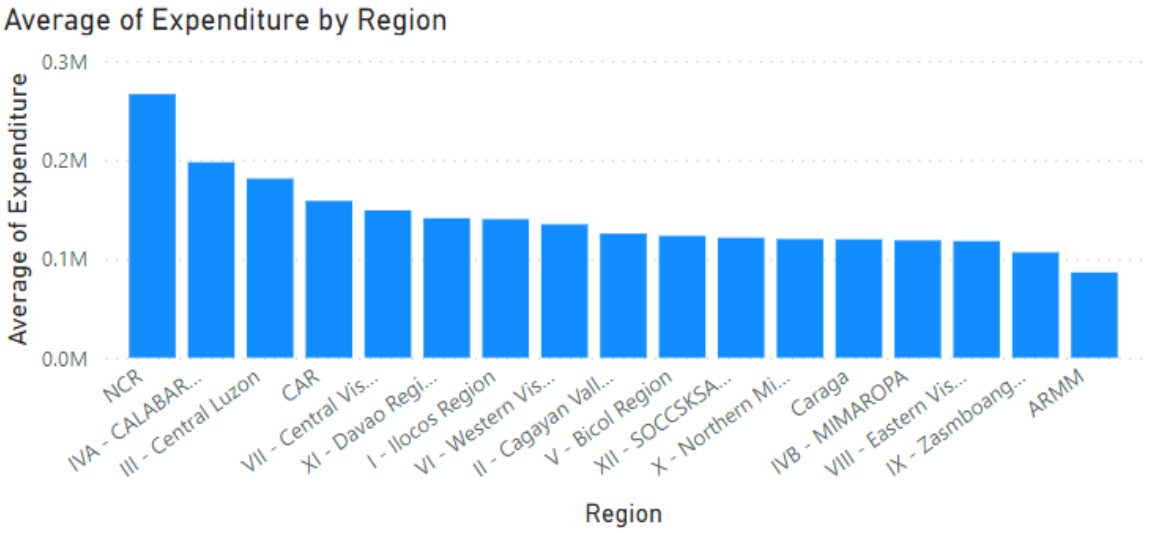
The data was further analyzed and the following observation was made (1) the NCR region has the highest average total household income with Region IV-A with the second highest.



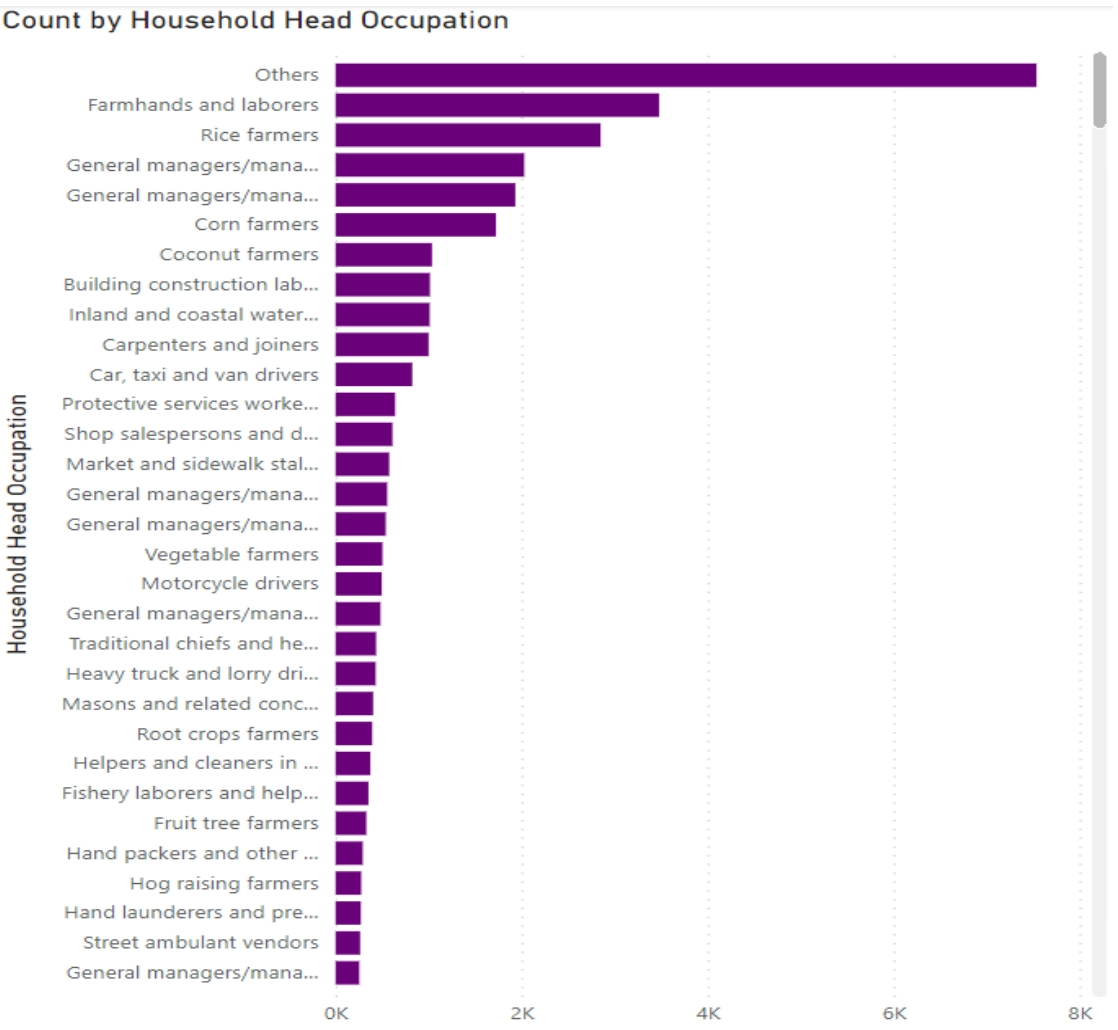
Most families have four family members with the highest number of family members in the dataset being 26.



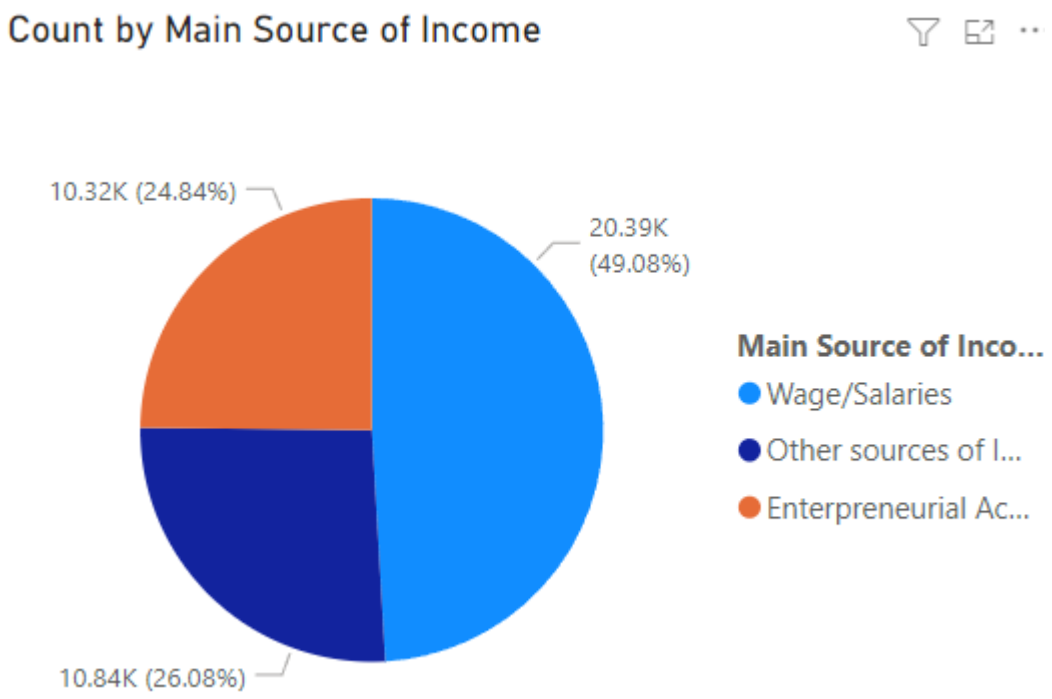
Average expenditure per region shows that in most regions people spend about half of their income on average with NCR still being the highest average expenditure followed by Region IV-A.



There are 379 unique household head occupations in the dataset with “Others” being the highest count followed by farmhands and laborers. There are single occupations in the dataset as well.



There are three main sources of income featured in the dataset. These are Wage/Salaries, Entrepreneurial Activities, and Other sources of Income. Wage/Salaries being the main source of income for 49% of the population. 26% goes to other sources of income, and 24.84% goes to entrepreneurial activities.



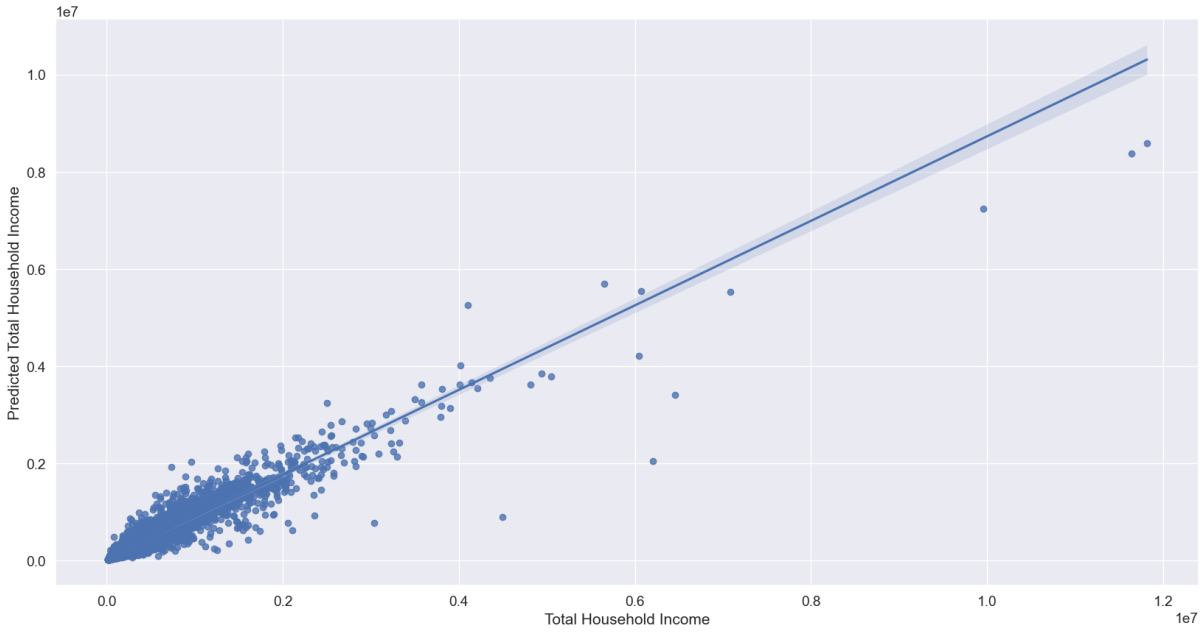
The chosen goal variable is total household income. This will be used to predict income classes. The potential predictors for this would be the top 3 correlated expenditures. The scatter plots show very little positive correlation between the goal variable. Hence, the choice for modeling would be random forest. This will allow classification and regression. Regression will only be used in this study. With the heavily skewed dataset and chaotic scatter plots, it is expected that the model would not be very accurate. It may be improved by adding more potential predictors or by utilizing the possessions or assets.

MODELING AND EVALUATION

The model chosen was random forest. The dataset was divided into two for training and testing. 80 percent of the dataset is for training while 20 percent is for testing. Sci-kit learn python library was used to develop the regression model. RandomForestRegressor was used. After the model was fitted with the training set. It was then tested. Using the top three expenditures, the result showed that the model predicted 69% of the data correctly. Using additional expenditures based on the most correlated variables, the model produced a score of 74%. This score may be improved by adding more predictors and possibly implementing classifiers.

Set	Number of data
Training set	33,235
Testing set	8,309

The graph shows where the predicted income lies in comparison to the actual income in a line. The closer the point to the line, the better the prediction performance of the model.



CONCLUSION AND RECOMMENDATION

This research aimed to predict the total household income in the Philippines using household expenditures as predictors. This helps determine the needed income for certain expenditures. The dataset used was taken from the Family Income and Expenditure Survey conducted by the Philippines Statistics Authority, and it contained 41, 544 rows of data.The data was processed, analyzed, and then used for modeling. The goal variable was the ‘Total Household Income’. The data was heavily skewed to the left, and the scatterplots did not show a clear relationship between the goal variable and the predictors. The predictors that were used for modeling were ‘Communication Expenditure', ‘Housing and water expenditure’, and Miscellaneous Goods and Services Expenditure’. A random forest regression model was used to predict total household income. The model yielded a 69% accuracy score. Using additional predictors namely 'Total Food Expenditure', 'Transportation Expenditure', 'Clothing, Footwear and Other Wear Expenditure', 'Imputed House

Rental Value', 'Meat Expenditure', the model yielded a 74% accuracy score. This may be improved by adding more predictors or using a classification model.

The researchers recommend thoroughly cleaning the data and removing most outliers. It is also possible to use total household monthly income instead of yearly. By using more predictors, the accuracy score may be improved. Data on possessions may also be used to predict household income or to classify the data in income classes. Possessions and household qualities may be supplementary to the current data set. These may help produce a higher accuracy score through classification.

REFERENCES

Benin, S., & Randriamamonjy, J. (2008). Estimating Household Income to Monitor and Evaluate Public Investment Programs in Sub-Saharan Africa. *International Food Policy Research Institute (IFPRI)*.

Berhanu, S. (1999). Econometric analysis of Household Expenditures. *Graduate Theses, Dissertations, and Problem Reports*. https://doi.org/10.33915/etd.3124

Briones, K. J. S., Lopez, J. S., Elumbre, R. J. L., & Angangco, T. G. (2021). Income, consumption, and poverty measurement in the Philippines. *Philippine Statistical Research and Training Institute*. Retrieved January 4, 2023, from https://mpra.ub.uni-muenchen.de/.

Francisco, E. de R., Aranha, F., Zambaldi, F., & Goldszmidt, R. (1970, January 1). *Electricity consumption as a predictor of household income: A spatial statistics approach*. SpringerLink. Retrieved January 4, 2023, from https://link.springer.com/chapter/10.1007/978-3-540-73414-7_17

Matkowski, M. (2021, April). *Prediction of individual income: A machine learning approach*. digitalcommons.bryant.edu. Retrieved January 4, 2023, from https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1038&context=honors_economics

Sri, Y. B., Sravani, Y., Surendra, Y. B., Rishitha, S., & Sobhana, M. (2021). Family expenditure and income analysis using machine learning algorithms. *2021 Second International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*. https://doi.org/10.1109/icstcee54422.2021.9708583

Villejo, S. J., Melendrez, M. J., Tan, D., & Enriquez, M. T. (2014, January). *Determinants of Income Class in Philippine Households: Evidence from the Family Income and Expenditure Survey 2009*. ResearchGate. Retrieved January 4, 2023, from https://www.researchgate.net/publication/274067118_Determinants_of_Income_Class_in_Philippine_Households_Evidence_from_the_Family_Income_and_Expenditure_Survey_2009

Wooden, M. (2004). The Effects of Wealth and Income on Subjective Well-Being and Ill-Being∗. *IZA*, *1*(1), 3-26. https://docs.iza.org/dp1032.pdf

APPENDIX

Github repository: https://github.com/evad-e/Business-proj.git

- Python Script
- Raw data
- Power BI report