

|  |  |  |
| --- | --- | --- |
|  | **PES UNIVERSITY, Bangalore**  (Established under Karnataka Act No. 16 of 2013) | **UE18CS203** |
| **B.Tech, Sem III**  **Session : Aug-Dec, 2019**  **UE18CS203 – INTRODUCTION TO DATA SCIENCE** | | |

**REPORT ON**

**EXPLORATORY ANALYSIS ON**

**FILIPINO FAMILY INCOME AND EXPENDITURE DATASET**

**SECTION :**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **#** | **SRN** | **Name** | **Contact No.** | **Email ID** | **Sign** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**ABOUT THE DATA SET**

The Family Income and Expenditure Survey (FIES), which is undertaken every three (3) years in the Phillipines, by the Philippine Statistics Authority (PSA) is aimed at providing data on family income and expenditure, including, among others, levels of consumption by item of expenditure, sources of income in cash, and related information affecting income and expenditure levels and patterns in the Philippines.

The dataset contains more than 40k observations and 60 variables which is primarily comprised of the household income and expenditures of that specific household.

The dataset describes the families by region, total household income, main source of income, family strength, family house floor area etc. which can help us to categorise the families for better analysis. It also lays out their expenditure in great detail by describing their food expenditure (subcategorised by rice, bread, meat) , housing expenditure, communication expenditure (includes number of cellphones, computers, TVs etc) and vehicles (cars, motorcycles etc.), all of which can help us to determine the key drivers of every field.

**ABSTRACT**

The task is to perform Exploratory Data Analysis on the Filipino Family Income and Expenditure Dataset, in order to gain valuable insights which can help to better understand the data, which in turn can be used to develop models to allow for prediction or future trends in the data.The analysis performed on the dataset is intended to observe the general trend in the income and expenditure, and the various factors which drive it. The families were divided into high income and middle income to allow for a more normalised and deeper understanding of the factors.

Using the data, we can derive how every region is doing financially, and the factors which derive it. We can also derive the how well the major drivers of income, such as educational level, type of income (Wages, Entrepreneurship etc) and region (which can determine their job availability), correlate in both the middle and high income social classes.

Thus we have depicted the levels of disparity in living conditions and spending patterns between each economic class, and their causes. Also an overall migration trend was noticed among growing income families.

**EXPLORATORY ANALYSIS**

As almost none of the variables in the dataset had missing or wrong values, no correction of data was performed. After plotting the boxplot of the total household income, the target variable for which we have tried to gain an understanding, we have chosen to classify all values below the upper whisker of the boxplot as middle income and all values above it (outliers) as high income. There were no values below the lower whisker of the boxplot, thus our classification allowed us to use all the values in the dataset for analysis. However, for some graphs, the extreme outliers (from the boxplot) in the high income class were ignored. The other variables were left intact.

Both the middle income and high income classes were plotted as a box plot by region. This, along with a bar graph showing the average income by region, showed which regions were financially well off and which regions were not. While the high income class did not vary much by region and had (almost) an equal median salary, the middle income class varied much by region. It was observed that the median middle class income tended to be higher in the densely populated metropolitan areas such the National Capital Region (NCR) which had a relatively high population, and tended to be lower in the agricultural and industrial regions such as the ARMM, which tended to have a lower population.

The pie chart depicting the Household Head Sex proportion depicted that most families had a male head. However the bar graph of the Houshold Head unemployment ratio, showed that female heads had a higher unemployment ratio (0.4) than the male heads (0.1). This shows that filipino families have a patriarchal structure, with most of the men being the income bringers for the family, while the women tend to be in charge of the household work.

A bar graph of average household income vs educational level showed that Law Programs (which are expensive and require many years of study) give the largest income, followed by other high level educational programs. Very low grade level correlated with a low average income, with No Grade Completed having the lowest income. Thus the educational level is a major driving factor of income.

Pie charts of sources of income for both the middle and high income classes showed that Wages was the dominant source of income in both classes. However it was noticed that proportion of enterpreneurs among high income families was much lower than that of middle income families. This suggested that the ease of doing business is not very good in the Phillipines. It also suggested that richer families tended to opt for the ‘safe’ option of wages / other sources of income rather than to risk their income on an entrepreneurship, thus giving the low ratio. One of the possible causes for this could be the fear of failure, which could ruin a person’s, and in turn, their families reputation. As most middle income families in the poorer regions are heavily dependent on selling agricultural produce for income, this could be a cause for the large entrepreneurship ratio among the middle income class.

Scatter plots of Alcohol Expenditure vs Tobacco Expenditure and Household Floor Area vs Household Income were plotted. It was observed that there was very little linear relationship in the former, with most of the larger values coming from the richer and highly populated regions. In the latter, it was observed that there was also very little linear relationship. However, it was observed that the there were very few households with a large house area among the high income class, while most of the larger houses belonged to the middle income class. Also it was observed that most the larger houses were present in the poorer regions while the richer families tended to stay in smaller houses in the richer regions, particularly in the NCR. This shows that as people tend to seek out smaller houses in richer regions rather than bigger houses in the poorer regions as their income grew. Causes of this could be the better job opportunities and better public facilities in the richer regions.

As there were many expenditure variables that were dependent on the Total Household Income, a correlation matrix was plotted in order to gain a better understanding of the target variable. A correlation matrix of the top ten variables most correlated with the Total Household Income were also plotted. It was shown that families regarded communication (which includes TVs, computers), housing, food and transportation expenditures as important as they were willing to spend more on them as their income grew, as depicted by the high correlation.

From the correlation matrix, it was observed that the Imputed House Rental Value was highly correlated with the Housing and Water Expenditure. A scatter plot was drawn to test the linear relationship and it was observed that the variables were linearly related. After passing the eye test, a linear regression model was chosen to predict the House Rental Value given the Housing and Water Expenditure. A best-fit line was plotted through the scatter plot and a few sample test values of Housing and Water Expenditure was given and the House Rental Value was predicted, and it was observed that the model tended to fit the data lower expenditure levels well, but predicted slightly lower rental values at higher expenditure levels.Thus, rental value tended to increase slightly faster as the housing and water expenditure increased. This showed that high income families (which tended to live in smaller houses in richer regions, mostly apartments) prefered well maintained homes and were willing to pay extra rent for it. However the middle income families tended to prefer more value for money while choosing their homes. Thus, when a landlord decides to rent out homes, the model can be used to predict the appropriate rental value from the housing expenditure depending on the type of family (middle income or high income) to rent out to.

**CONCLUSION**

Thus we have depicted the levels of disparity in living conditions and spending patterns between each economic class, and have also determined the root causes that put families in each economic class. We have also determined the overall migration trend, in which growing income families tend to move to richer regions, even if some of their expenditures increase.