**Dataset Description**

# Filipino Family Income and Expenditure

## **TASK**

One dataset will be provided to each team. The dataset can be found in the folder corresponding to the team number.

The team’s task is to understand the dataset completely, and perform exploratory data analysis on it. This would entail using knowledge acquired from the class sessions to plot visualizations which would help understand the dataset better, and derive insights from it. These visualizations and basic insights are a must for the project. However, the focus of this project is to augment theoretical learning of the class with invaluable hands-on knowledge.

This project is also meant to increase awareness and encourage exploration of fields and topics which are overlapping with the field of Data Science. Thus, a portion of the project is meant to test the team’s willingness to learn and explore about new topics and techniques. Take note, however, that these might be topics which are not part of the Data Science course, but nonetheless will impart a great deal of knowledge.

To aid the thinking process, there is a section titled ‘Possible tracks for insights’ which is meant to provide an example of a possible direction to begin thinking in. The questions provided in that section are NOT a checklist of things to be done, as the basic visualizations need to be implemented, even though they are not mentioned in the section. It could also be the case that the question provided in this section does not provide any insight. In such cases, simply consider it as a direction to be thinking in, and feel free to skip the results of the question in the report. Marks will be awarded for any and all techniques not part of the course which have been implemented.

## **DESCRIPTION**

A dataset is a collection of data. In a dataset, each row is a particular instance or record of an event, and each column corresponds to one variable or feature.

Inside this data set is some selected variables from the latest Family Income and Expenditure Survey (FIES) in the Philippines. It contains more than 40k observations and 60 variables which is primarily comprised of the household income and expenditures of that specific household

**Take note**, however, that the data may have to be cleaned, with a number of erroneous or blank values, with a number of outliers. The dataset will be a CSV file or a number of them. Your task is to clean the dataset, and perform analysis on it, plot relevant visualizations and to gain some meaningful insight into the data using the Data Science skills you have acquired. If you deem any columns to be irrelevant to your analysis, you can discard them. However, you will have to report clearly why this column was discarded.

## **COLUMN EXPLANATION**

In this section, a short description of the meaning of each column can be found.

1. Total Household Income
2. Region
3. Total Food Expenditure
4. Main Source of Income
5. Agricultural Household indicator
6. Bread and Cereals Expenditure
7. Total Rice Expenditure
8. Meat Expenditure
9. Total Fish and marine products Expenditure
10. Fruit Expenditure
11. Vegetables Expenditure
12. Restaurant and hotels Expenditure
13. Alcoholic Beverages Expenditure
14. Tobacco Expenditure
15. Clothing, Footwear and Other Wear Expenditure
16. Housing and water Expenditure
17. Imputed House Rental Value
18. Medical Care Expenditure
19. Transportation Expenditure
20. Communication Expenditure
21. Education Expenditure
22. Miscellaneous Goods and Services Expenditure
23. Special Occasions Expenditure
24. Crop Farming and Gardening expenses
25. Total Income from Entrepreneurial Acitivites
26. Household Head Sex
27. Household Head Age
28. Household Head Marital Status
29. Household Head Highest Grade Completed
30. Household Head Job or Business Indicator
31. Household Head Occupation
32. Household Head Class of Worker
33. Type of Household
34. Total Number of Family members
35. Members with age less than 5 year old
36. Members with age 5 - 17 years old
37. Total number of family members employed
38. Type of Building/House
39. Type of Roof
40. Type of Walls
41. House Floor Area
42. House Age
43. Number of bedrooms
44. Tenure Status
45. Toilet Facilities
46. Electricity
47. Main Source of Water Supply
48. Number of Television
49. Number of CD/VCD/DVD
50. Number of Component/Stereo set
51. Number of Refrigerator/Freezer
52. Number of Washing Machine
53. Number of Airconditioner
54. Number of Car, Jeep, Van
55. Number of Landline/wireless telephones
56. Number of Cellular phone
57. Number of Personal Computer
58. Number of Stove with Oven/Gas Range
59. Number of Motorized Banca
60. Number of Motorcycle/Tricycle

## **POSSIBLE TRACKS FOR INSIGHTS**

In this section, a small number of possible insights by manipulating and analysing the data are presented. This is certainly **NOT** a checklist of things to do in your project, but rather aims to provide a generic example of the direction to begin thinking in. It could certainly be the case that more valuable insights may be possible, which are not listed here.

Is there a correlation between Household Head Highest Grade Completed and total income?

What factors are total income correlated to? What can you conclude?

Which region is the richest?

What do the rich ( > median average income) spend most of their money on? How about the poor?

Is there a correlation between alcoholic beverage expenditure and tobacco expenditure?

## **POTENTIALLY USEFUL LINKS**

In this section, you will find a number of links, which can be used to brush up your knowledge or help resolve issues.

### Python Tutorial

### <https://www.learnpython.org/>

### Conda Cheatsheet

<https://docs.conda.io/projects/conda/en/4.6.0/_downloads/52a95608c49671267e40c689e0bc00ca/conda-cheatsheet.pdf>

### Pandas Documentation

<https://pandas.pydata.org/pandas-docs/stable/>

### Pandas Quickstart (Short Tutorial)

<https://pandas.pydata.org/pandas-docs/stable/getting_started/10min.html>

### NumPy Quickstart (Short Tutorial)

<https://docs.scipy.org/doc/numpy/user/quickstart.html>

### Seaborn Documentation

<https://seaborn.pydata.org/introduction.html>

### Seaborn In-Depth Tutorial

<https://seaborn.pydata.org/tutorial.html#tutorial>

### Matplotlib Tutorial

<https://matplotlib.org/3.1.1/tutorials/index.html>

### Scikit-Learn Tutorial

<https://scikit-learn.org/stable/tutorial/index.html>