Data Analysis on Filipino Family Income and Expenditure

By: Johann Sebastian E. Catalla, BSCS-1

An output for the course CP102: Computer Programming 2

This dataset was sourced through Kaggle from the Philippine Statistics Authority's (PSA) Family Income and Expenditure Survey (FIES) nationwide. The survey, which is undertaken every three (3) years, is a nationwide sample survey of households with detailed information, including, among others, levels of consumption by item of expenditures, sources of income in cash, and related information affecting income and expenditure patterns of households 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, for this analysis, I will be using Python **pandas**, **numpy**, and **Kaggle** libraries.

```
import pandas as pd
import numpy as np
import kagglehub
path = kagglehub.dataset_download("ananyaananya/philippine-expenditure-income")
filename = "Family Income and Expenditure.csv"
full_path = os.path.join(path, filename)
print("Dataset downloaded to:", full_path)
```

```
Dataset downloaded to: C:\Users\Johann Sebastian Catalla\.cache\kagglehub\datasets\ananyaananya\philippine-expenditure-income\versions\1\Family Income and Expenditure.csv
```

```
df = pd.read_csv(full_path, encoding='latin-1')
df.head()
```

Total Household Income	Region	Total Food Expenditure	Main Source of Income	Agricultural Household indicator	Bread and Cereals Expenditure	Total Rice Expenditure	Total meat expenditure	Total fish and other seafood expenditure	Total Fruit Expenditure	Number of Refrigerator/Freezer	Number of Washing Machine	Number of Aircon/Fan	Number of Television
480000	13 NCR	117696	Wage/Salaries	1	16320	16320	16320	16320	16320	1	1	1	1
198000	13 NCR	107766	Wage/Salaries	0	17029	12888	17034	15072	2910	0	1	0	0
267000	13 NCR	91009	Wage/Salaries	0	16920	22001	7992	2208	1728	1	1	0	0
107000	13 NCR	78195	Wage/Salaries	0	16020	20820	10914	10812	336	1	1	0	0
189000	13 NCR	78624	Wage/Salaries	0	16020	20757	16020	11208	1200	1	1	0	0

410220	170775	05 BATAANGAS	48875	Entrepreneurial Activities	1	25675	27540	9870	3139	1632	1	0	0
415680	137550	05 BATAANGAS	21157	Entrepreneurial Activities	1	2501	1373	1956	4396	1440	1	0	0
415000	132979	05 BATAANGAS	4800	Entrepreneurial Activities	1	19440	27108	480	4776	1032	1	0	0
415440	120900	05 BATAANGAS	35196	Entrepreneurial Activities	1	20896	28003	2559	10728	2823	1	0	0
415680	120900	05 BATAANGAS	35196	Entrepreneurial Activities	1	41480	41395	1965	7755	2862	1	0	0

4708 rows × 60 columns

QUESTIONS

1. What is the Average Yearly Food Expenditure and Average Income per region? Which region has the highest and lowest?
2. What is the ratio of Food Expenditure to income per region? Which region has the highest and lowest?

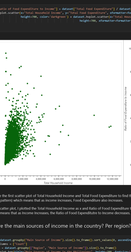
```
# Group by region and calculate the average yearly food expenditure
df_grouped = df.groupby('Region')['Total Food Expenditure', 'Total Household Income']
print("Grouped data:")
print(df_grouped.head())
```

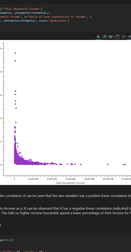
```
Grouped data:
                Total Food Expenditure  Total Household Income
Region
1 ILOCOS REGION                  1...  ...
2 CAGAYAN VALLEY                 1...  ...
3 CENTRAL LUZON                  1...  ...
4 CALABARZON                     1...  ...
5 MIMAROPA                       1...  ...
```

```
# Calculate the mean for each group
mean_values = df_grouped.mean()
print("Mean values:")
print(mean_values.head())
```

```
Mean values:
                Total Food Expenditure  Total Household Income
Region
1 ILOCOS REGION                 ...  ...
2 CAGAYAN VALLEY                ...  ...
3 CENTRAL LUZON                 ...  ...
4 CALABARZON                    ...  ...
5 MIMAROPA                      ...  ...
```

```
# Create a bar chart
import matplotlib.pyplot as plt
plt.figure(figure=(10, 6))
mean_values['Total Food Expenditure'].plot(kind='bar', color='skyblue')
plt.title('Average Yearly Food Expenditure')
plt.xlabel('Region')
plt.ylabel('Average Yearly Food Expenditure')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```





It is evident that the National Capital Region (NCR) has the highest Average Household Income (AHI) and the Highest Average Food Expenditure (AFE). Moreover, the Autonomous Region of Muslim Mindanao (ARMM) has the lowest Average Household Income (AHI) while Northern Mindanao the lowest Average Food Expenditure (AFE).

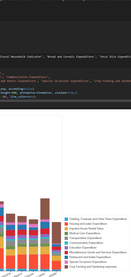
In terms of Ratio of Food Expenditure to Income, CAVO (MIMAROPA) has the highest ratio with 0.5, meaning 50% of their income goes to food. On the other hand, the region with the lowest ratio is NCR, with 0.37 or 37%. Despite having the highest income and highest food expenditure, NCR has the lowest ratio of food to income, which implies that their other expenses are marginally more prioritized compared to other regions.

It can be observed that in lower income regions are more likely to allot a higher percentage of their income to food expenditures. Similarly, Higher income regions are more likely to have a lower percentage of their income allotted for food expenditures.

3. Scatter plot of Household Income and Total Food Expenditure in the Philippines
4. Scatter plot of Household Income and Ratio of Food Expenditure to Income

```
dataset['Ratio of Food Expenditure to Income'] = dataset['Total Food Expenditure'] / dataset['Total Household Income']
dataset.dropna(subset=['Total Food Expenditure', 'Total Household Income', 'Ratio of Food Expenditure to Income'], inplace=True)
plt.figure(figure=(10, 6))
sns.scatterplot(x='Total Food Expenditure', y='Total Household Income', data=dataset, alpha=0.7)
plt.title('Household Income vs Total Food Expenditure')
plt.xlabel('Total Food Expenditure')
plt.ylabel('Total Household Income')
plt.show()
```

```
plt.figure(figure=(10, 6))
sns.scatterplot(x='Ratio of Food Expenditure to Income', y='Total Household Income', data=dataset, alpha=0.7)
plt.title('Household Income vs Ratio of Food Expenditure to Income')
plt.xlabel('Ratio of Food Expenditure to Income')
plt.ylabel('Total Household Income')
plt.show
```



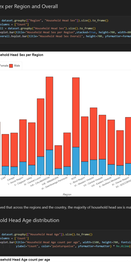
We can observe the first scatter plot of Total Household Income and Total Food Expenditure to find the correlation. It can be seen that the two variables has a positive linear correlation (indicated by the upward trend) which means that as income increases, Total Expenditure also increases.

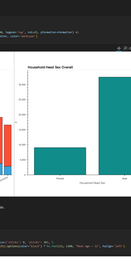
In the second scatter plot, I plotted the Total Household Income as y and Ratio of Food Expenditure to Income as x. It can be observed that it has a negative linear correlation (indicated by the lower-inclined trend) which means that as income increases, the Ratio of Food Expenditure to Income decreases. We, with an higher income households spend a lower percentage of their income for food.

5. What are the main sources of income in the country? Per region?

```
counts = dataset.groupby('Main Source of Income').size().to_frame().sort_values(ascending=False)
print("Income counts:")
print(counts.head(3))
counts.plot(kind='bar', color=['blue', 'orange', 'green'], figure=(10, 6))
plt.title('Main Sources of Income in the Philippines')
plt.xlabel('Main Source of Income')
plt.ylabel('Number of Households')
plt.xticks(rotation=0, ha='center')
plt.tight_layout()
plt.show()
```

```
Income counts:
                      0
Main Source of Income
Wage/Salaries      2...
Other sources of Income  ...
Entrepreneurial Activities  ...
```





In the first bar graph, it can be observed that Wage/Salaries is the main source of income Filipino Household with Entrepreneurial Activities being the lowest.

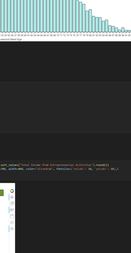
We can also see that, in the second bar graph, only ARMM has Entrepreneurial Activities being the higher source of income. The rest of the regions has Wage/Salaries as their main source of income.

6. How much does each region spend on other expenses?

```
keys = dataset.columns
keys.tolist()
for k in keys:
    print(k)
    break
```

```
['Total Household Income', 'Region', 'Total Food Expenditure', 'Main Source of Income', 'Agricultural Household indicator', 'Bread and Cereals Expenditure', 'Total Rice Expenditure', 'Total Meat Expenditure', 'Total Fish and Marine Products Expenditure', 'Total Fruit Expenditure', 'Total Vegetable Expenditure', 'Restaurant and Hotels Expenditure', 'Transport and Communication Expenditure', 'Clothing, Footwear and Other Wear Expenditure', 'Housing and Water Expenditure', 'Imputed House Rent Value', 'Medical Care Expenditure', 'Transportation Expenditure', 'Communication Expenditure', 'Education Expenditure', 'Miscellaneous Goods and Services Expenditure', 'Personal Care Items Expenditure', 'Special Occasions Expenditure', 'Crop Farming and Gardening expenses', 'Total Income from Entrepreneurial Activities', 'Total Number of Family members', 'Total number of family members employed', 'House Floor Area', 'House Roof Quality', 'Number of Bedrooms', 'Electricity', 'Water Supply', 'Toilet Facilities', 'Number of Refrigerator/Freezer', 'Number of Washing Machine', 'Number of Aircon/Fan', 'Number of Television', 'Number of Cellular Phones', 'Number of Light Bulbs', 'Number of Cars or Calesa', 'Number of Motorcycle/Tricycle', 'Number of Computer/Laptop with Internet Connection', 'Number of Computer/Laptop without Internet Connection', 'Number of Stereo with Speaker/s', 'Number of Gas Range', 'Number of Electric Fan', 'Number of Electric Rice Cooker', 'Number of Electric Flat Iron', 'Number of Electric Oven/Toaster', 'Number of Electric Refrigerator', 'Number of Electric Stove', 'Number of Water Dispenser', 'Number of Microwave Oven', 'Number of Washing Machine', 'Number of Aircon/Fan', 'Number of Television']
```

```
other_exp = ['Clothing, Footwear and Other Wear Expenditure', 'Housing and water Expenditure',
             'Imputed House Rent Value', 'Medical Care Expenditure', 'Transportation Expenditure', 'Communication Expenditure',
             'Education Expenditure', 'Miscellaneous Goods and Services Expenditure', 'Personal and Care Items Expenditure', 'Special Occasions Expenditure', 'Crop Farming and Gardening expenses']
expenses_counter = dataset.groupby('Region')[other_exp].mean().reset_index().sort_values(ascending=True)
expenses_counter.melt(id_vars=['Region'], var_name='Other expenses for graph per Region', value_name='Average').\
                 pivot(index='Region', columns='Other expenses for graph per Region').\
                 sort_values('Average', ascending=False).\
                 plot(figure=(10, 6), stacked=True, legend=True, xlabel='', ylabel='Average', rot=90, ylim=(0, 50000))
plt.legend(bbox_to_anchor=(1.0, 1.0), loc='upper left', ncol=1, fontsize=8, title='Other expenses for graph per Region')
plt.tight_layout()
plt.show()
```

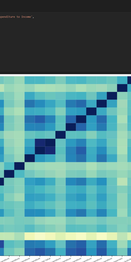


The following conclusions can be made:

1. NCR has the highest housing and water, house rental, transportation, clothing and footwear, communication, education, restaurant and hotels, and miscellaneous goods and service expenditures.
2. NCR has the lowest crop farming/gardening expenses.
3. CAR has the lowest house rental, communication, and travel expenditures for the rest of the categories.
4. Cagayan Valley and Caraga Administrative Region (CAR) has the second and third highest expenditure on crop farming/gardening respectively.
5. CALABARZON holds the highest expenditure for special occasions and medical care.
6. MA has the highest travel expenditure for other regions.

7. Head Sex per Region and Overall

```
head_sex = dataset.groupby('Region', 'Household Head Sex').size().to_frame()
head_sex.columns = ['Count']
head_sex = head_sex.reset_index(level=1).reset_index()
head_sex.hvplot.barh(x='Household Head Sex', y='Count', groupby=[], height=500, width=400, legend=False, rot=45, stacked=True, alpha=0.7)
plt.title('Household Head Sex Per Region')
plt.xlabel('Count')
plt.ylabel('Household Head Sex')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



It can be observed that across the regions and the country, the majority of household head sex is male.

8. Household Head Age distribution

```
head_age = dataset.groupby('Household Head Age').size().to_frame()
head_age.columns = ['Count']
head_age.hvplot.bar(x='Household Head Age', count_age=[], width=500, height=400, fontsize={'xlabel': 8, 'xticks': 8}, \
                    label='Count', ylabel='Count', alpha=0.7, xlabel='Household Age', ylabel='Count', rot=90, ylim=(0, 1200), title='Head Age', ylim=(0, 1200))
```

It can be observed that the bar graph forms a bell curve.

The mean age is 51.28 which is sorted by the following code.

Mean Age

```
# Calculate the header columns to take out column
sex_data = head_age.iloc[:,0:1]
sex_data.columns = ['Count']
sex_data = sex_data.reset
sex_data.sort_values(ascending=False, axis=1)
sex_data.columns = ['Household Head Age', 'Count']
sex_data.head()
mean = (sex_data['Household Head Age'] * sex_data['Count'])
mean = mean.sum() / sex_data['Count'].sum()
print("Mean Age: ", mean)
```

```
Mean Age:  51.28398305085
```

9. Regions sorted by Total Income from Entrepreneurial Activities

```
entrep = dataset.groupby('Region')['Total Income from Entrepreneurial Activities'].mean().sort_values('Total Income from Entrepreneurial Activities').reset()
entrep.hvplot.barh(x='Region', y='Total Income from Entrepreneurial Activities', height=500, width=400, color='Crimson', ylabel='Total Income from Entrepreneurial Activities', xlabel='Region', \
                   alpha=0.7, xlim=(0, 20000))
```


It can be concluded that ARMM has the highest Total Income from Entrepreneurial Activities while Bicol Region has the lowest.

10. Correlation of Income and Expenditures

```
correlation = dataset['Total Household Income', 'Total Food Expenditure', 'Ratio of Food Expenditure to Income',
                      'Total Expenditure', 'Total Rice Expenditure', 'Total Meat Expenditure',
                      'Total Vegetable Expenditure', 'Total Fish and Marine Products Expenditure',
                      'Total Fruit Expenditure', 'Total Income from Entrepreneurial Activities',
                      'Total Number of Family members', 'Total number of family members employed',
                      'House Floor Area', 'Number of Bedrooms', 'Electricity', 'Water Supply',
                      'Toilet Facilities', 'Number of Refrigerator/Freezer', 'Number of Washing Machine',
                      'Number of Aircon/Fan', 'Number of Television', 'Number of Cellular Phones',
                      'Number of Light Bulbs', 'Number of Cars or Calesa', 'Number of Motorcycle/Tricycle',
                      'Number of Computer/Laptop with Internet Connection', 'Number of Computer/Laptop without Internet Connection',
                      'Number of Stereo with Speaker/s', 'Number of Gas Range', 'Number of Electric Fan',
                      'Number of Electric Rice Cooker', 'Number of Electric Flat Iron', 'Number of Electric Oven/Toaster',
                      'Number of Electric Refrigerator', 'Number of Electric Stove', 'Number of Water Dispenser',
                      'Number of Microwave Oven', 'Number of Washing Machine', 'Number of Aircon/Fan', 'Number of Television']
correlation.hvplot.heatmap(height=1000, height=800, rot=45, xlim=(0, 1))
```


The heatmap above plots the correlation between the variables of Expenses together with the Total Household Income and Food Percentage to Income. It can be observed that Ratio of Food Expenditure to Income and Total Household Income has a moderately strong negative correlation (0.45), which confirmed our findings earlier. Furthermore, Total Food Expenditure and Total Household Income has a strong positive correlation (0.6), which also confirmed our findings.

The following conclusions can be made:

1. Total Household Income has a significant effect to Total Food Expenditure. Higher income households are more likely to spend more on food.
2. Total Household Income has a significant effect to Ratio of Food Expenditure to Income. Higher income households are more likely to have a lower percentage of their income alloted for food expenditures. Similarly, lower income households are more likely to allot a higher percentage of their income for food expenditures.
3. Total Household Income has a significant effect to Other Expenses. Higher income households are more likely to spend more on expenses other than food.

The following conclusions can be made:

Stop Scroll this content in the selected area