

Group_01_Project2_demo

Group_01

Introduction

Data come from the FIES (Family Income and Expenditure Survey) recorded in the Philippines. The survey, which is undertaken every three years, is aimed at providing data on family income and expenditure. The data obtained from this survey are from different regions across the Philippines. This report will focus on one individual area, the Cordillera Administrative Region and so region has been removed from the dataset as it will not be informative as an explanatory variable.

The report will investigate which household related variables influence the number of people living in a household. The data used consists of 1725 observations of ten variables, two of which are categorical and the remaining are numerical.

Exploratory Data Analysis

Figure 1 shows the distribution of the response variable: Number of members in a household (variable name “Total.number.of.family.members”). The modal response is 4 members and the distribution is right-skewed,

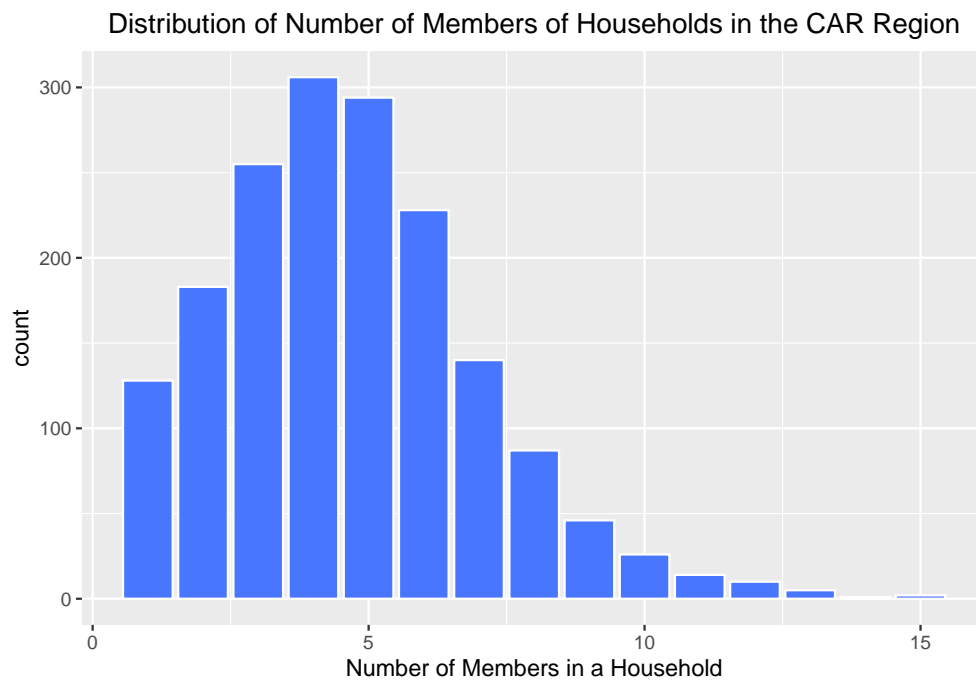


Figure 1: Distribution of Response Variable

Table 1: Summary statistics of numerical variables

Variable	Missing	Complete	Mean	SD	Min.	1st Q.	Median	3rd Q.	Max.
Total.Number.of.Family.members	0	1	4.67	2.33	1.00	3.00	4.00	6.00	15.00
Total.Household.Income	0	1	26.95	27.46	1.20	11.86	18.86	32.83	604.29
Total.Food.Expenditure	0	1	8.04	4.12	0.68	5.19	7.36	9.85	32.77
Household.Head.Age	0	1	52.23	14.52	17.00	41.00	52.00	63.00	99.00
House.Floor.Area	0	1	90.92	99.20	5.00	32.00	54.00	102.00	900.00
House.Age	0	1	22.98	15.32	0.00	12.00	20.00	31.00	100.00
Number.of.bedrooms	0	1	2.26	1.44	0.00	1.00	2.00	3.00	9.00
Electricity	0	1	0.93	0.26	0.00	1.00	1.00	1.00	1.00

Table 2: Correlation of all variables.

	Total.Number.of.Family.members	Total.Household.Income	Total.Food.Expenditure	Household.Head.Age	House.Floor.Area	House.Age	Number.of.bedrooms	Electricity
Total.Number.of.Family.members	1.000	0.192	0.469	-0.065	-0.014	-0.070	0.072	0.092
Total.Household.Income	0.192	1.000	0.611	0.063	0.234	0.025	0.441	0.149
Total.Food.Expenditure	0.469	0.611	1.000	-0.052	0.124	0.007	0.356	0.199
Household.Head.Age	-0.065	0.063	-0.052	1.000	0.091	0.218	0.154	-0.013
House.Floor.Area	-0.014	0.234	0.124	0.091	1.000	0.074	0.374	0.107
House.Age	-0.070	0.025	0.007	0.218	0.074	1.000	0.123	0.085
Number.of.bedrooms	0.072	0.441	0.356	0.154	0.374	0.123	1.000	0.214
Electricity	0.092	0.149	0.199	-0.013	0.107	0.085	0.214	1.000

The summary below shows the count data for each level of the response variable and the percentage of total households in the region in each group.

Table 1 shows summary data for all the numerical variables. There is no missing data within these variables and so no values will need to be imputed for the analysis in the report. The response variable, total number of family members in a household, ranges from 1 to 15, with the middle 50% of number of family members falling between 3 and 6 also an average number of family members of 4.67. There appear to be possible outliers at the maximum values of Total Household Income and House Floor Area. The total household income is range from 11988 to 6042860 Philippine Peso. The middle 50% of total household income is between 118565 and 328335, with an income of 269540.48 peso on average. The third variable is total food expenditure, in Phillipine peso, which is the range of 6781 to 327724, with the middle 50% lies between 51922 and 98493. Then, the household head's age is range from 17 to 99 years, with the middle 50% falling between 41 and 63 years of age. Next, the house floor area is range from 5 to 900 square metres. The central 50% of the variable house floor area is between 32 and 102 with an average area of 90.92. The sixth explanatory variable is the house (building) age and it ranges in value from 0 to 100, with the middle 50% falling between 12 and 31. The number of bedrooms in the house ranges from 0 to 9 with a mean average number of bedrooms of 2.26 per household. Finally, we may look at the binary variable electricity, which denotes whether the property has electricity access or not, the average score of electricity is 0.93, which means 93% household have electricity.

The correlation coefficient between all numerical variables are shown in Table 2. There is a moderate positive correlation (0.611) between the total household income and the household food expenditure. Additionally there is a slight positive correlation between the total household income and the number of bedrooms in the household (0.441) and the number of family members and total food expenditure (0.469). The other variables are all weakly correlated. The correlation coefficient between household Head's age, house floor area, house age and total number of family members are negative, which shows the rise of those three variables will lead to a reduction in the expected number of family members in a household.

Table 3 shows the summaries of the two categorical variables. Single family households make up approximately two-thirds of the survey responses in this region and only 0.5% (8) of responses came from households formed from non-related individuals. Of the 1725 households, less than a quarter (21.4%) had a female head of household.

Table 3: Summary of Categorical Explanatory Variables

Household.Head.Sex	n	percent
Female	369	21.4%
Male	1356	78.6%
Total	1725	100.0%

Type.of.Household	n	percent
Extended Family	569	33.0%
Single Family	1148	66.6%
Two or More Nonrelated Persons/Members	8	0.5%
Total	1725	100.0%

The pairs plot in Figure 2 is colour coded to illustrate any differences between the distributions of the quantitative variables when the head of household sex is included as a factor. The plots suggest the sex of the head of household may impact the number of family members in the household and the age of the head of the household.

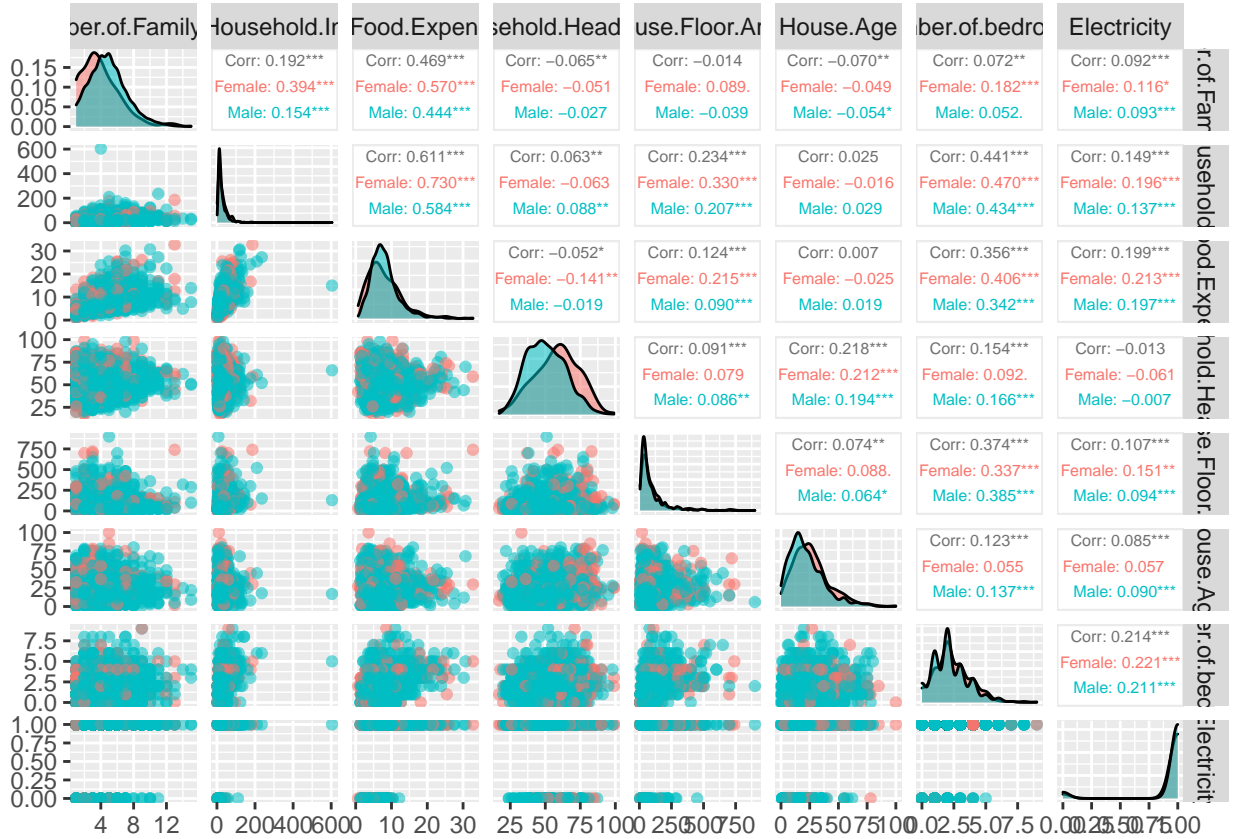


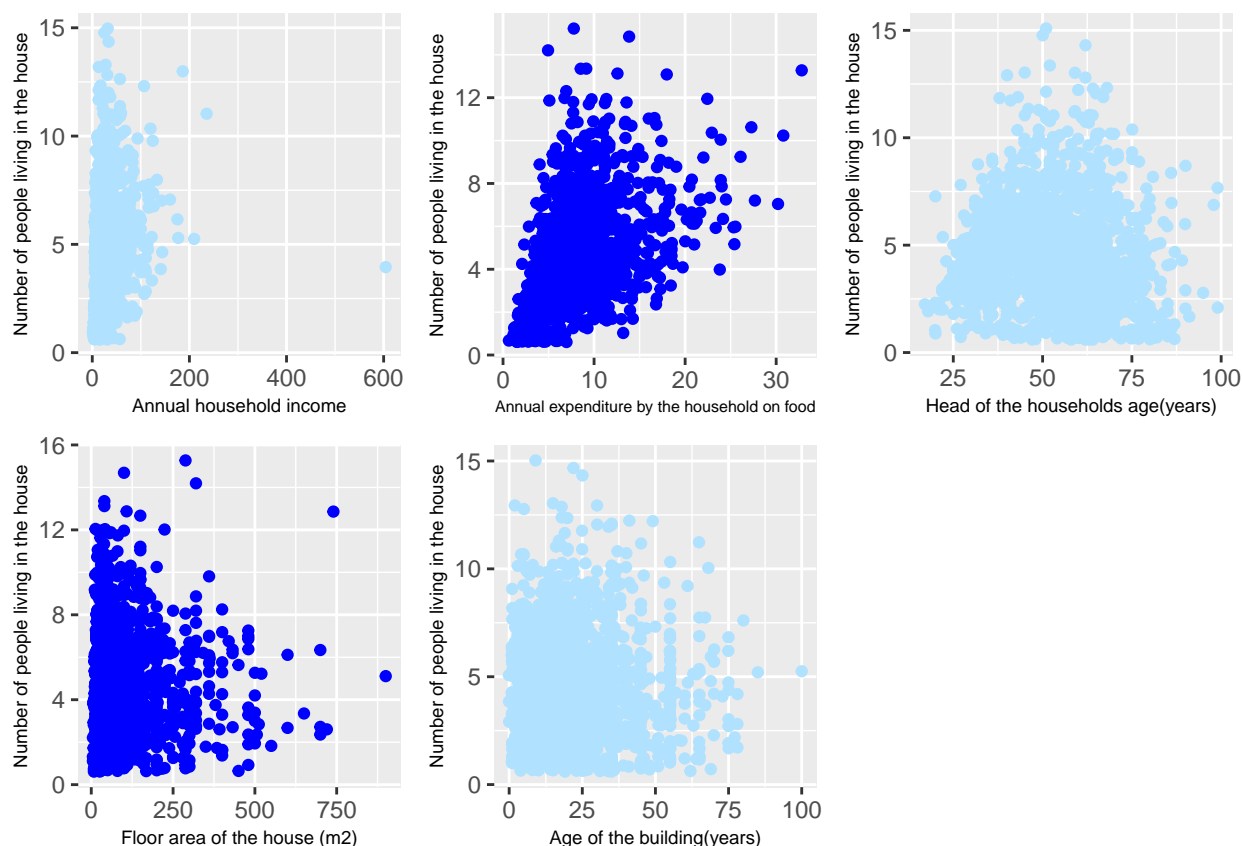
Figure 2: Paired plots of the variables

Analysis of Relationships between Covariates

The relationship between independent and dependent variables

We will explore whether nine variables related to households in the dataset have an impact on the number of people living in the house (Total.Number.of.Family.members). These nine variables are Annual household income, Annual expenditure by the household on food, Head of the households sex, Head of the households age, Relationship between the group of people living in the house, Floor area of the house, Age of the building, Number of bedrooms in the house and House electrification situation.

First let's look at the scatterplot of the response variable versus the numerical variable.



Gender & Age

As highlighted by the pairs plot, there appears to be a relationship between the sex and age of the head of the household.

The minimum and maximum ages of household heads do not appear to differ greatly according to the individuals' sex, however they do differ at the 25th, 50th and 75th percentiles with male heads of households being consistently younger than their female counterparts. The standard deviation is also greater for the female group, but the substantially smaller group size for females may contribute to this larger variation.

The boxplot in Figure 4 illustrates the previously summarised data. The boxplot identifies the two oldest male head of households as outliers (shown by the points above the whisker), however within the context of the data and when compared to the ages of female head of household boxplot, these ages do not appear unreasonable or unrealistic.

Table 4: Summary statistics on the age of household heads by sex.

Household.Head.Sex	n	Mean	St.Dev	Min	Q1	Median	Q3	Max
Female	369	58.23	15.69	17	47	59	69	99
Male	1356	50.59	13.74	20	40	49	61	98

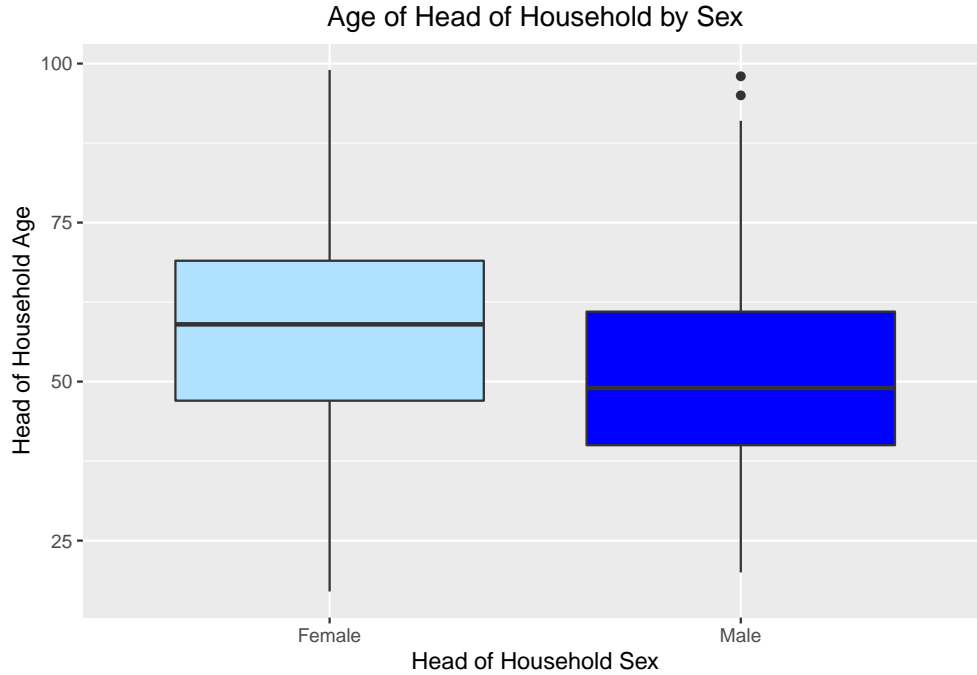


Figure 3: Boxplots of Head of Household Age stratified by Sex

The following Mann-Whitney U-test shows that there is a statistically significant difference in the median ages of male and female head of households at a 5% level.

Wilcoxon rank sum test with continuity correction

```
data: data.gender$Household.Head.Age by data.gender$Household.Head.Sex
W = 324284, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Household Income and Food Expenditure

Figure 5 shows a boxplot of household incomes suggests a heavily skewed distribution with many outliers at the upper end of the distribution.

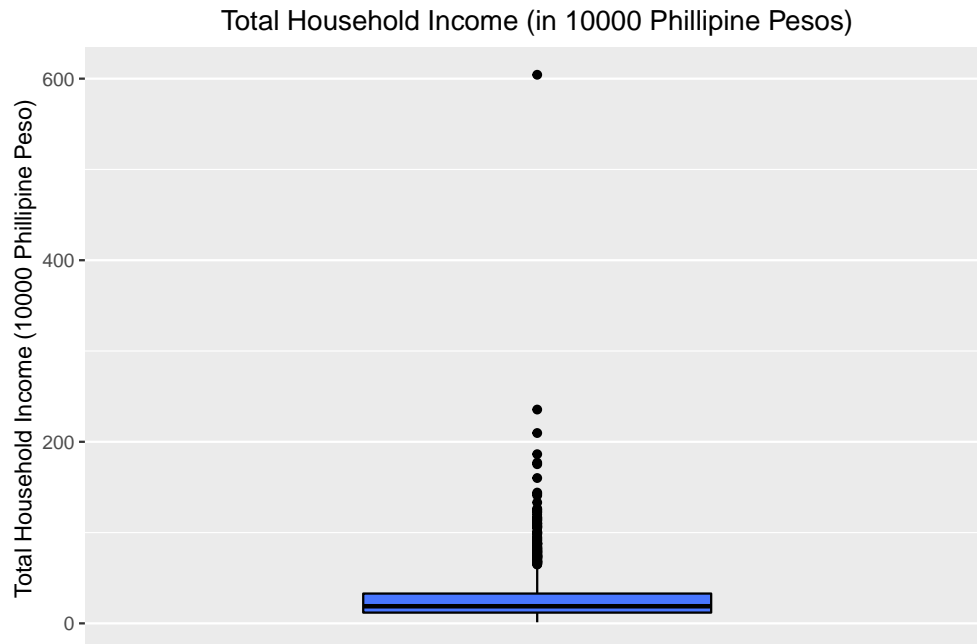


Figure 4: Household Incomes in 10000 Phillipine Pesos.

The following boxplot in Figure 6 shows the log transformed household income and shows there are still several outliers following the transformation.

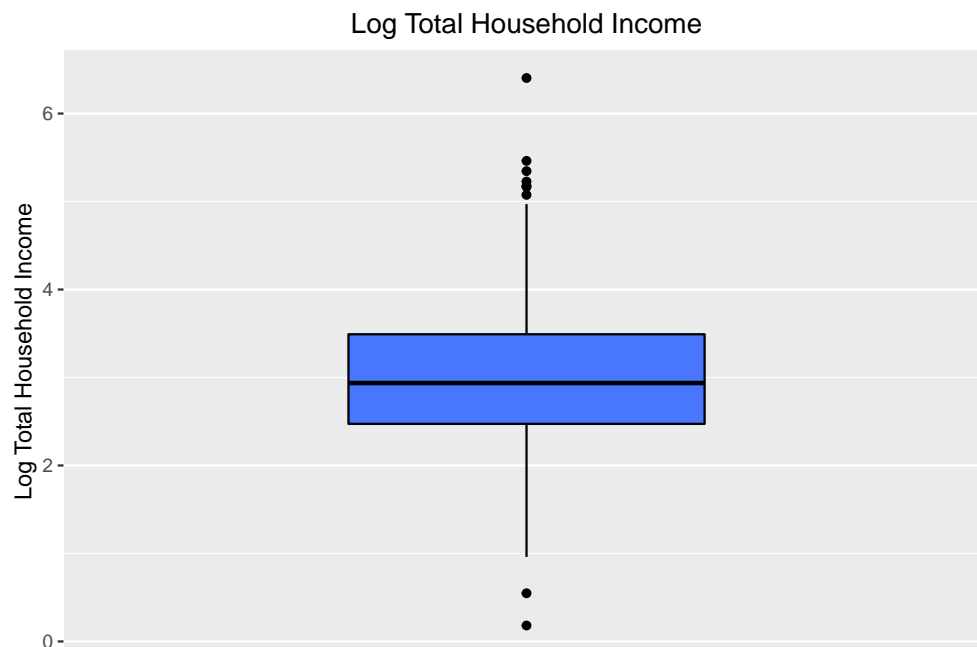


Figure 5: Boxplot of log transformed Incomes.

The scatterplot in Figure 7 of household income against total food expenditure suggests a positive correlation

but the fitted model may be being heavily influenced by the extreme values, particularly by one extreme point.

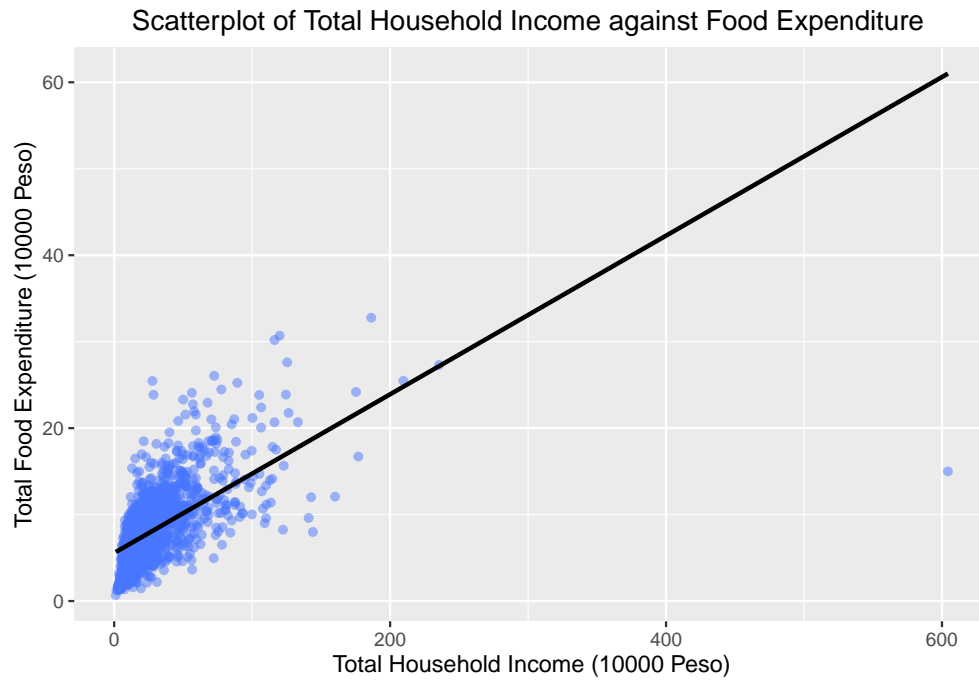


Figure 6: Scatterplot of Income against Food Expenditure.

Figure 7 again highlights a possible outlier in terms of income, this could be a data entry error or just an outlier at the maximum. Removing this observation from the data set and plotting provides the following scatter diagram in Figure 8. This plot reconfirms the suggested positive correlation, but there is still an imbalance in the amount of data available at different levels of income. For example, most data is available for incomes between 0 and 750000 peso, but far fewer data points occur above this income level.

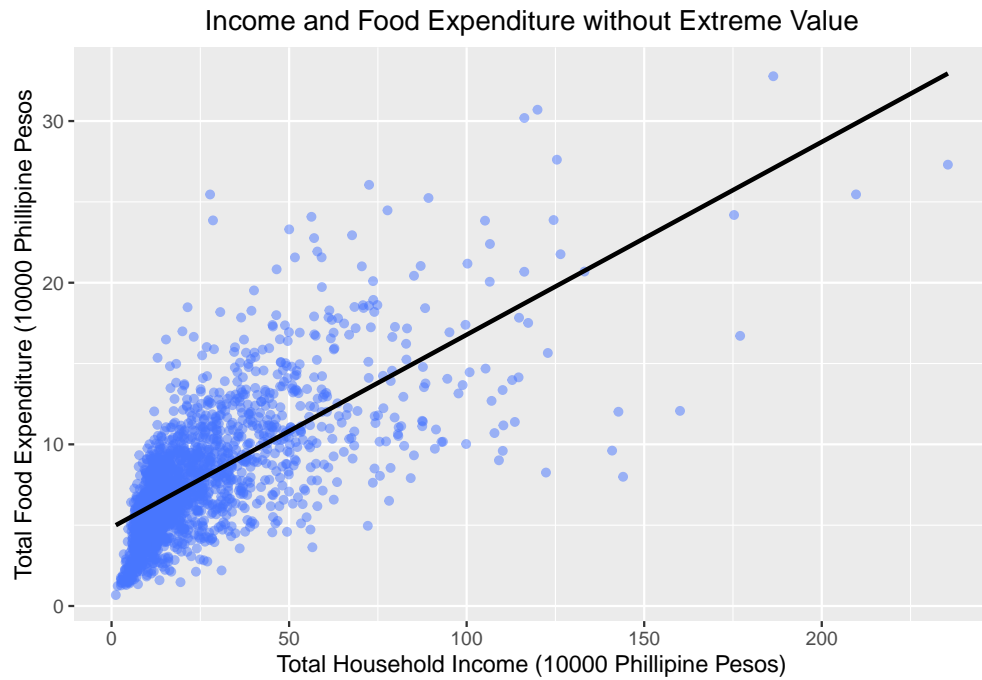


Figure 7: Scatterplot of Income and Food Expenditure with extreme value removed.

Number of Family Members & Head of Household Sex

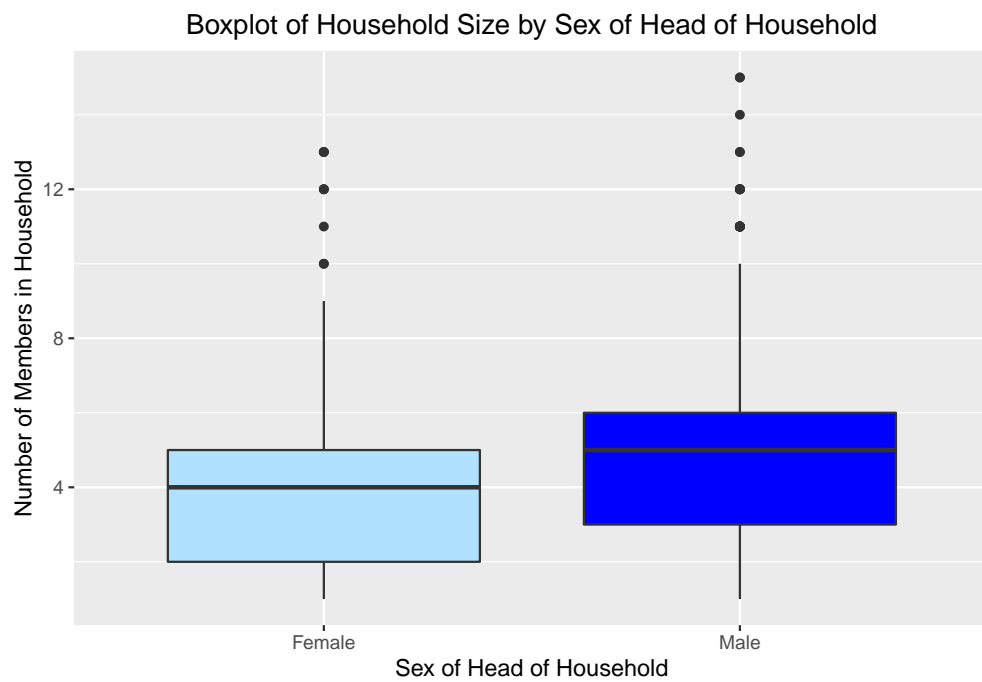


Figure 8: Number of Members in Household by Sex of Head of Household

Hence we can see from Figure 9 that households with a male head appear to have a greater number of family members on average than those with a female head, as the male group has larger values for the first and third quartiles and the median. However there is overlap between the two groups central IQR and so the distributions may not be significantly different.

Figure 9 shows that an extended family household or one formed by non-related individuals is more likely to have a female head, whereas a larger proportion of single family households have male heads.

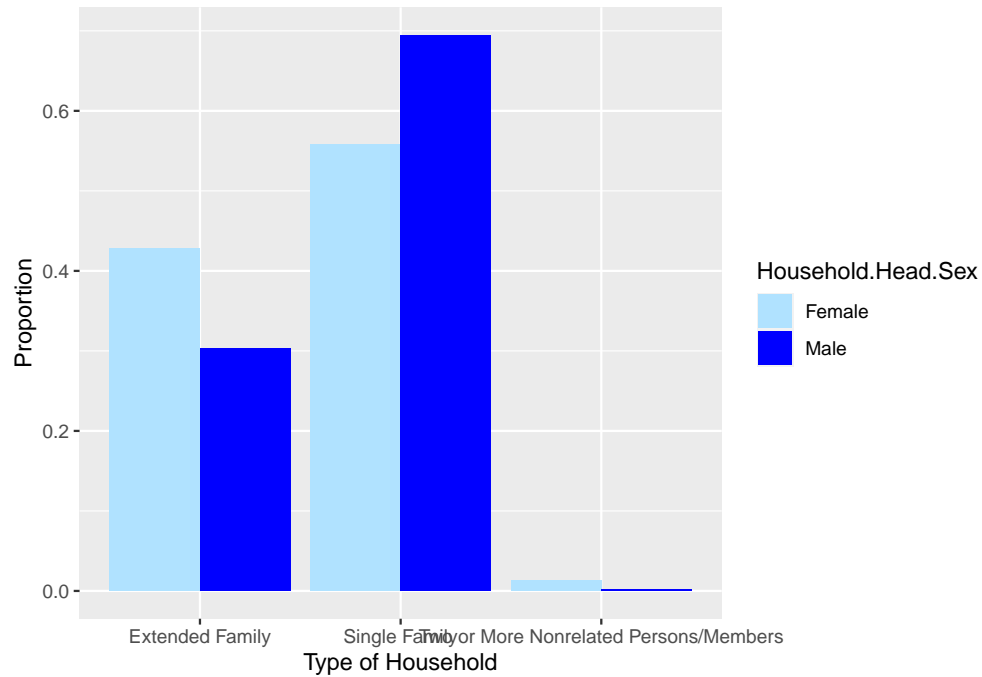


Figure 9: Barplot of household head's sex by type of household

Exploratory Model Analysis

GLM Model Exploration

Prior to exploring any models, the outlier for Total Household Income and corresponding measurements for the other variables from this individual are removed.

The following code identifies which explanatory variables would be included to produce the best models of different sizes, in this instance the maximum number of variables specified is ten. The output suggests the first predictor to be included is the total food expenditure in 10000 Phillipine pesos, and the last to be included is the binary variable Electricity that identifies if a household has electricity. Comparing each of the ten models produced by BIC, CP and adjusted R^2 criteria is inconclusive as each implies a different model is best.

Subset selection object

```
Call: regsubsets.formula(Total.Number.of.Family.members ~ ., data = data,
  nvmax = 10)
```

10 Variables (and intercept)

	Forced in	Forced out
Total.Household.Income	FALSE	FALSE
Total.Food.Expenditure	FALSE	FALSE
Household.Head.SexMale	FALSE	FALSE
Household.Head.Age	FALSE	FALSE
Type.of.HouseholdSingle Family	FALSE	FALSE
Type.of.HouseholdTwo or More Nonrelated Persons/Members	FALSE	FALSE
House.Floor.Area	FALSE	FALSE
House.Age	FALSE	FALSE
Number.of.bedrooms	FALSE	FALSE
Electricity	FALSE	FALSE

1 subsets of each size up to 10

Selection Algorithm: exhaustive

	Total.Household.Income	Total.Food.Expenditure	Household.Head.SexMale
1 (1)	" "	"*"	" "
2 (1)	" "	"*"	" "
3 (1)	" "	"*"	"*"
4 (1)	"*"	"*"	"*"
5 (1)	"*"	"*"	"*"
6 (1)	"*"	"*"	"*"
7 (1)	"*"	"*"	"*"
8 (1)	"*"	"*"	"*"
9 (1)	"*"	"*"	"*"
10 (1)	"*"	"*"	"*"

	Household.Head.Age	Type.of.HouseholdSingle Family
1 (1)	" "	" "
2 (1)	" "	"*"
3 (1)	" "	"*"
4 (1)	" "	"*"
5 (1)	" "	"*"
6 (1)	" "	"*"
7 (1)	"*"	"*"
8 (1)	"*"	"*"
9 (1)	"*"	"*"
10 (1)	"*"	"*"

Type.of.HouseholdTwo or More Nonrelated Persons/Members

1	(1)	" "			
2	(1)	" "			
3	(1)	" "			
4	(1)	" "			
5	(1)	" "			
6	(1)	" "			
7	(1)	" "			
8	(1)	" "			
9	(1)	"*"			
10	(1)	"*"			
		House.Floor.Area	House.Age	Number.of.bedrooms	Electricity
1	(1)	" "	" "	" "	" "
2	(1)	" "	" "	" "	" "
3	(1)	" "	" "	" "	" "
4	(1)	" "	" "	" "	" "
5	(1)	" "	"*"	" "	" "
6	(1)	" "	"*"	"*"	" "
7	(1)	" "	"*"	"*"	" "
8	(1)	"*"	"*"	"*"	" "
9	(1)	"*"	"*"	"*"	" "
10	(1)	"*"	"*"	"*"	"*"
Adj.R2	CP	BIC			
9	8	6			

Table 5: Mean and Variation of Response Variable

Mean	4.669373
Variance	5.446049

Formal Model Analysis

Poisson Regression model

The response variable of the Total Number of Family Members (or members of the household) can be viewed as a count and therefore a Poisson Regression model is considered. For a Poisson model to be suitable, the mean and variance should be equal and so these assumptions are checked first.

From the table 5, The variation of total number of family members is only marginally larger than the mean of total number of family members, thus, the possibility of over-dispersion in our model is not a significant issue.

Call:

```
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Age + House.Floor.Area +
    House.Age + Number.of.bedrooms + Electricity + Household.Head.Sex +
    Type.of.Household, family = poisson(link = "log"), data = data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.7749	-0.6993	-0.1044	0.4989	3.7501

Coefficients:

	Estimate	Std. Error	
(Intercept)	1.4254422	0.0796515	
Total.Household.Income	-0.0022045	0.0006612	
Total.Food.Expenditure	0.0500316	0.0035528	
Household.Head.Age	-0.0025205	0.0008707	
House.Floor.Area	-0.0001932	0.0001281	
House.Age	-0.0023168	0.0007735	
Number.of.bedrooms	-0.0145830	0.0095600	
Electricity	0.0276347	0.0475502	
Household.Head.SexMale	0.2202770	0.0297157	
Type.of.HouseholdSingle Family	-0.3481835	0.0248020	
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-0.1444455	0.1598841	
	z value	Pr(> z)	
(Intercept)	17.896	< 2e-16	***
Total.Household.Income	-3.334	0.000856	***
Total.Food.Expenditure	14.082	< 2e-16	***
Household.Head.Age	-2.895	0.003793	**
House.Floor.Area	-1.509	0.131385	
House.Age	-2.995	0.002744	**
Number.of.bedrooms	-1.525	0.127155	
Electricity	0.581	0.561126	
Household.Head.SexMale	7.413	1.24e-13	***
Type.of.HouseholdSingle Family	-14.039	< 2e-16	***
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-0.903	0.366293	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2024.3 on 1723 degrees of freedom
Residual deviance: 1330.4 on 1713 degrees of freedom
AIC: 7011.6

Number of Fisher Scoring iterations: 4

The poisson model fitted with all possible covariates concludes there are six statistically significant predictors at the 5% level. These are the total household income and food expenditure, the age and gender of the head of the household, the age of the house and if it is a single family household. Table 5 shows the estimates and the lower and upper bounds of the 95% confidence intervals for the regression parameters. The rows containing significant predictors, and so where the confidence intervals do not include 0, are highlighted.

\begin{table}

\caption{Estimates and the corresponding 95% Confidence Intervals, with significant predictors highlighted.}

	Lower Bound	Estimate	Upper Bound
(Intercept)	1.2687606	1.4254422	1.5810043
Total.Household.Income	-0.0035110	-0.0022045	-0.0009192
Total.Food.Expenditure	0.0430507	0.0500316	0.0569775
Household.Head.Age	-0.0042279	-0.0025205	-0.0008148
House.Floor.Area	-0.0004469	-0.0001932	0.0000552
House.Age	-0.0038392	-0.0023168	-0.0008069
Number.of.bedrooms	-0.0333633	-0.0145830	0.0041111
Electricity	-0.0644803	0.0276347	0.1219537
Household.Head.SexMale	0.1623556	0.2202770	0.2788479
Type.of.HouseholdSingle Family	-0.3967514	-0.3481835	-0.2995260
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-0.4743426	-0.1444455	0.1540719

\end{table}

We refit the model to include just the previously identified significant covariates and again evaluated the 95% confidence intervals for the estimated parameters, these values can be seen in Table 6. The intercept term of 1.436 is simply a positional constant due to the context of the variables. The negative coefficient of Total Household Income shows that for every additional 10000 peso, the number of household members is expected to decrease by 0.002. The coefficient of Total Food Expenditure suggests that for an increase of 10000 peso in spending, there is an expected 0.048 more members in the household. The coefficients of Head of Household age and the Age of the Building are both negative (-0.003 and -0.002 respectively) showing that an older head of the household or older building is linked to fewer members in a household. A

Single Family household is expected to have 0.350 fewer members than the baseline category of an extended family household, and households with a male head will have 0.222 members more than their female counterparts.

Call:

```
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Age + House.Age +
    Household.Head.Sex + Type.of.Household, family = poisson(link = "log"),
    data = data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.7325	-0.7095	-0.1012	0.5090	3.7690

Coefficients:

	Estimate	Std. Error
(Intercept)	1.4313832	0.0672809
Total.Household.Income	-0.0028211	0.0006146
Total.Food.Expenditure	0.0503772	0.0035264
Household.Head.Age	-0.0027721	0.0008627
House.Age	-0.0024807	0.0007682
Household.Head.SexMale	0.2205681	0.0297155
Type.of.HouseholdSingle Family	-0.3484617	0.0247448
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-0.1365104	0.1598583
	z value	Pr(> z)
(Intercept)	21.275	< 2e-16 ***
Total.Household.Income	-4.590	4.44e-06 ***
Total.Food.Expenditure	14.286	< 2e-16 ***
Household.Head.Age	-3.213	0.00131 **
House.Age	-3.229	0.00124 **
Household.Head.SexMale	7.423	1.15e-13 ***
Type.of.HouseholdSingle Family	-14.082	< 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-0.854	0.39313

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2024.3 on 1723 degrees of freedom
Residual deviance: 1336.7 on 1716 degrees of freedom
AIC: 7012

Number of Fisher Scoring iterations: 4

\begin{table}

\caption{Estimates of regression parameters and the corresponding 95% Confidence Intervals}

	Lower Bound	Estimate	Upper Bound
(Intercept)	1.2992842	1.4313832	1.5630233
Total.Household.Income	-0.0040352	-0.0028211	-0.0016258
Total.Food.Expenditure	0.0434480	0.0503772	0.0572714
Household.Head.Age	-0.0044639	-0.0027721	-0.0010820
House.Age	-0.0039924	-0.0024807	-0.0009811
Household.Head.SexMale	0.1626469	0.2205681	0.2791383
Type.of.HouseholdSingle Family	-0.3969151	-0.3484617	-0.2999140
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-0.4663624	-0.1365104	0.1619512

\end{table}

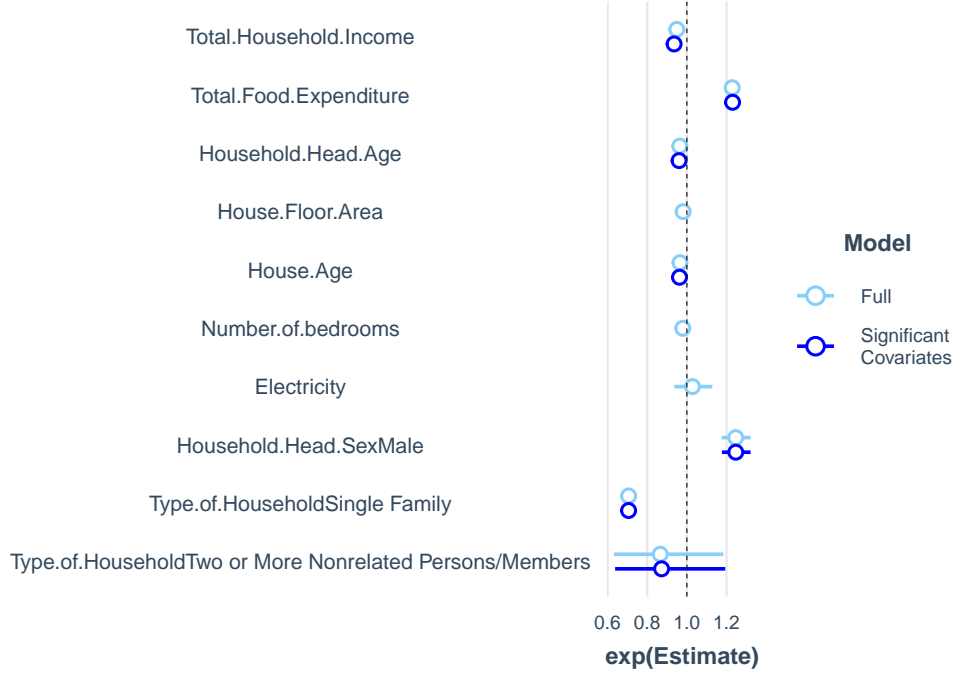


Figure 10: Summary of Coefficients for each fitted Poisson Model

Table 6: Comparison of Fitted Poisson Models

Model	AIC	BIC
Full Model	7011.64	7071.61
Significant Factors Model	7012.00	7055.62

From table8, we got the AIC value of two model do not have significant different, however, the BIC value of significant factors model is lower than the full model. Thus, we decided to choose the significant factors model.

The poisson regression model that will be fitted to the data is as follows:

$$\log(\tilde{\lambda}_i) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \hat{\beta}_{\mathbf{M}} \cdot \mathbb{I}_{\mathbf{M}}(i) + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2),$$

where

- $\log(\tilde{\lambda})$ is the Total Number of Family Members of the i^{th} individual;
- x_1 is the Total Household Income in pesos;
- x_2 is the Total Food Expenditure in pesos;
- x_3 is the Household Head's Age in years;
- x_4 is the House's Age in years;
- α is the intercept and positions the best-fitting plane in 3D space;
- β_1 is the coefficient for the first explanatory variable x_1 ;
- β_2 is the coefficient for the second explanatory variable x_2 ;
- β_3 is the coefficient for the second explanatory variable x_3 ;
- β_4 is the coefficient for the second explanatory variable x_4 ;
- $\hat{\beta}_{\mathbf{M}}$ i
- ϵ_i is the i^{th} random error component; and

- $\mathbb{I}_F(i)$ is an indicator function such that

$$\mathbb{I}_M(i) = \begin{cases} 1 & \text{if the household head of } i\text{th observation is Male,} \\ 0 & \text{Otherwise.} \end{cases}$$

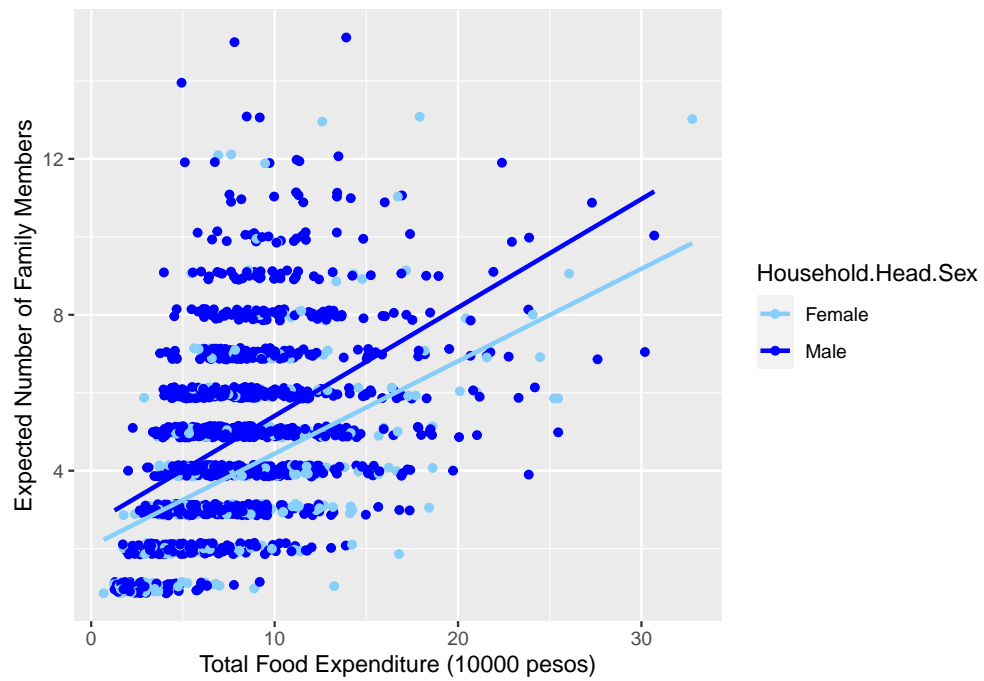


Figure 11: Predicted Numbers of Household Members



Figure 12: Scatterplot Numbers of Household Members and number of family members

Conclusions

In conclusion,
