Data Analysis Skills

Group 03

Calum Lawson, Savitha Sundaresan, Wenyu Qu, Yimeng Zhang, Yuezhu Guo

Contents

List of Figures

List of Tables

Introduction

Understanding factors that could potentially influence the number of people living in a household is essential for researchers and policymakers to gain insights into the social and economic conditions of households in a country. In this study, we will analyze the household dataset from the FIES (Family Income and Expenditure Survey) of the Philippines.

We will examine which household-related variables influence the number of people living in a household, including total household income, total food expenditure, household head's sex and age, type of household, the total number of family members, house floor area, house age, number of bedrooms, and electricity. Since the "Region" is the same (i.e., X-Northern Mindanao) throughout the given dataset, we will drop the "Region" variable from the data frame. Using a Generalized Linear Model (GLM), we will identify which of these variables have a statistically significant relationship with the number of people living in a household.

Variables

List of Variable are given in table 1

Table 1: Description of Variables

Variable	Description
Total.Household.Income	Annual household income (in Philippine peso)
Region	The region, X - Northern Mindanao of the Philippines
Total.Food.Expenditure	Annual expenditure by the household on food (in Philippine peso)
Household.Head.Sex	Head of the households sex
Household.Head.Age	Head of the households age (in years)
Type.of.Household	Relationship between the group of people living in the house
Total.Number.of.Family.members	Number of people living in the house
House.Floor.Area	Floor area of the house (in m^2)
House.Age	Age of the building (in years)
Number.of.bedrooms	Number of bedrooms in the house
Electricity	Does the house have electricity? (1=Yes, 0=No)

Since we have the same region throughout the dataset, we will consider dropping the Region from the dataframe.

Exploratory Data Analysis

Summary Statistics

Summary is given in Table 2:

Table 2: Summary

Total.Household.Income	Total.Food.Expenditure	Household.Head.Sex
Min. : 16238	Min. : 3704	Length:1887
1st Qu.: 85545	1st Qu.: 38311	Class :character
Median : 131806	Median : 54594	Mode :character
Mean : 214058	Mean : 64113	NA
3rd Qu.: 249176	3rd Qu.: 77068	NA
Max. :2598050	Max. :363572	NA

Household.Head.Age	Type.of.Household	Total.Number.of.Family.members
Min. :15.00	Length:1887	Min. : 1.000
1st Qu.:41.00	Class :character	1st Qu.: 3.000
Median :51.00	Mode :character	Median : 4.000
Mean :51.52	NA	Mean : 4.677
3rd Qu.:61.00	NA	3rd Qu.: 6.000
Max. :95.00	NA	Max. :16.000

House.Floor.Area	House.Age	Number.of.bedrooms	Electricity
Min. : 10.00	Min. : 0.0	Min. :0.000	No : 257
1st Qu.: 30.00	1st Qu.:10.0	1st Qu.:1.000	Yes:1630
Median : 50.00	Median :16.0	Median :2.000	NA
Mean : 59.81	Mean :19.5	Mean :1.945	NA
3rd Qu.: 80.00	3rd Qu.:26.0	3rd Qu.:2.000	NA
Max. :600.00	Max. :95.0	Max. :8.000	NA

Plots

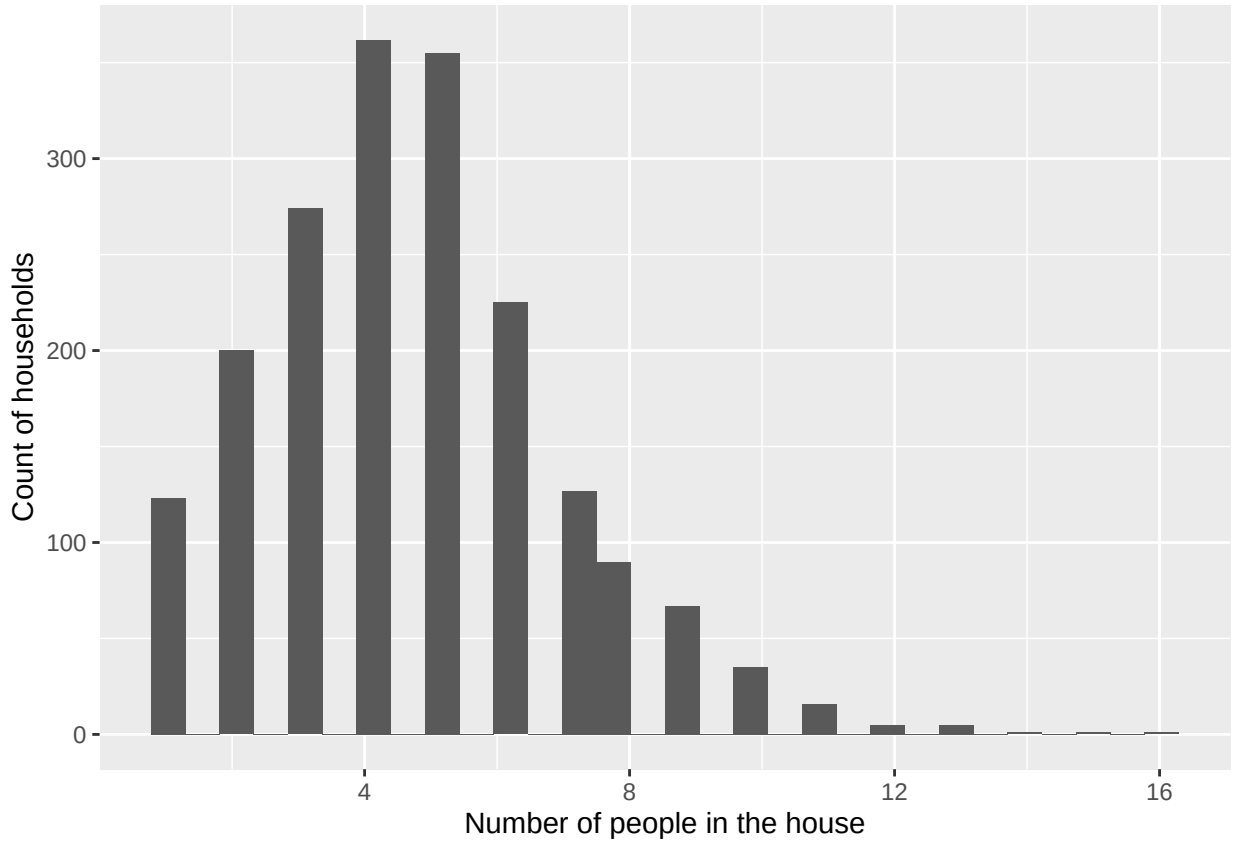

Figure 1: Histogram

From Figure 1, the histogram seems to be right skewed, suggesting that our variable under consideration, the number of people in a household is not normally distributed. Also, since the number of people is a count variable, Poisson regression will be a suitable fit.

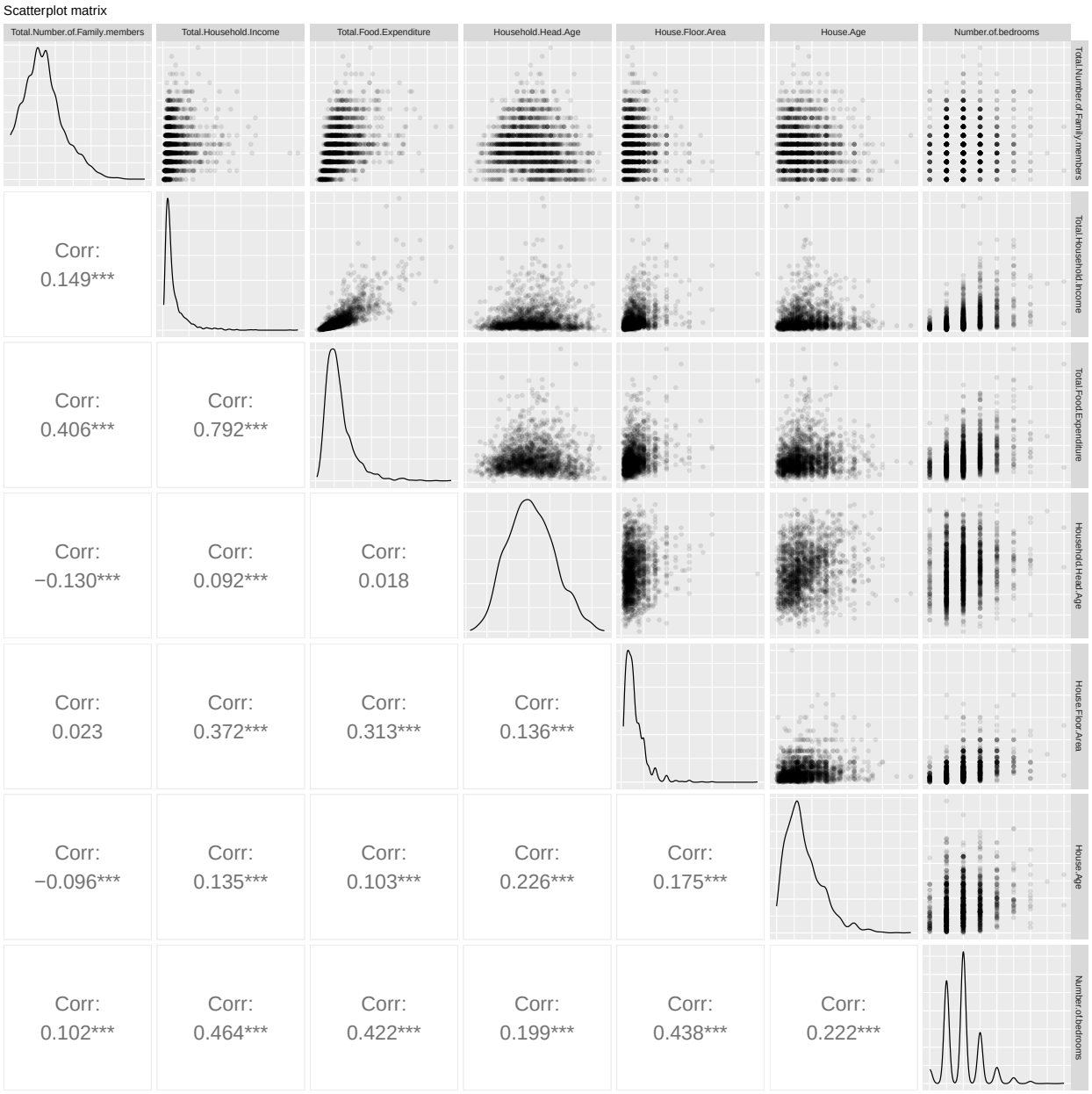

Figure 2: Scatterplot matrix

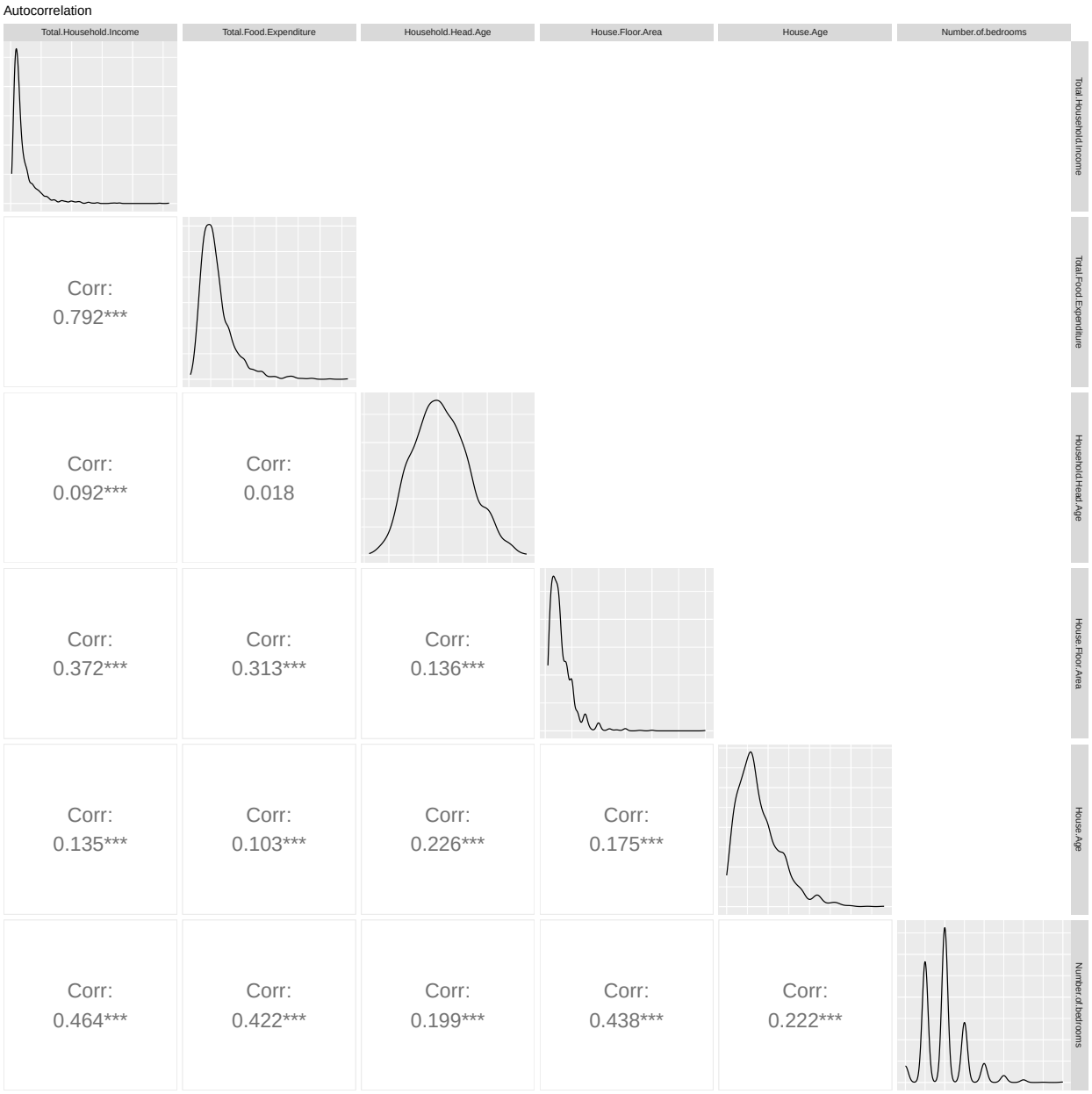

Figure 3: Autocorr

From Figure 3, there seems to be a strong correlation between Total Income and Expenditure on Food. We might consider dropping one of these variables from the model.

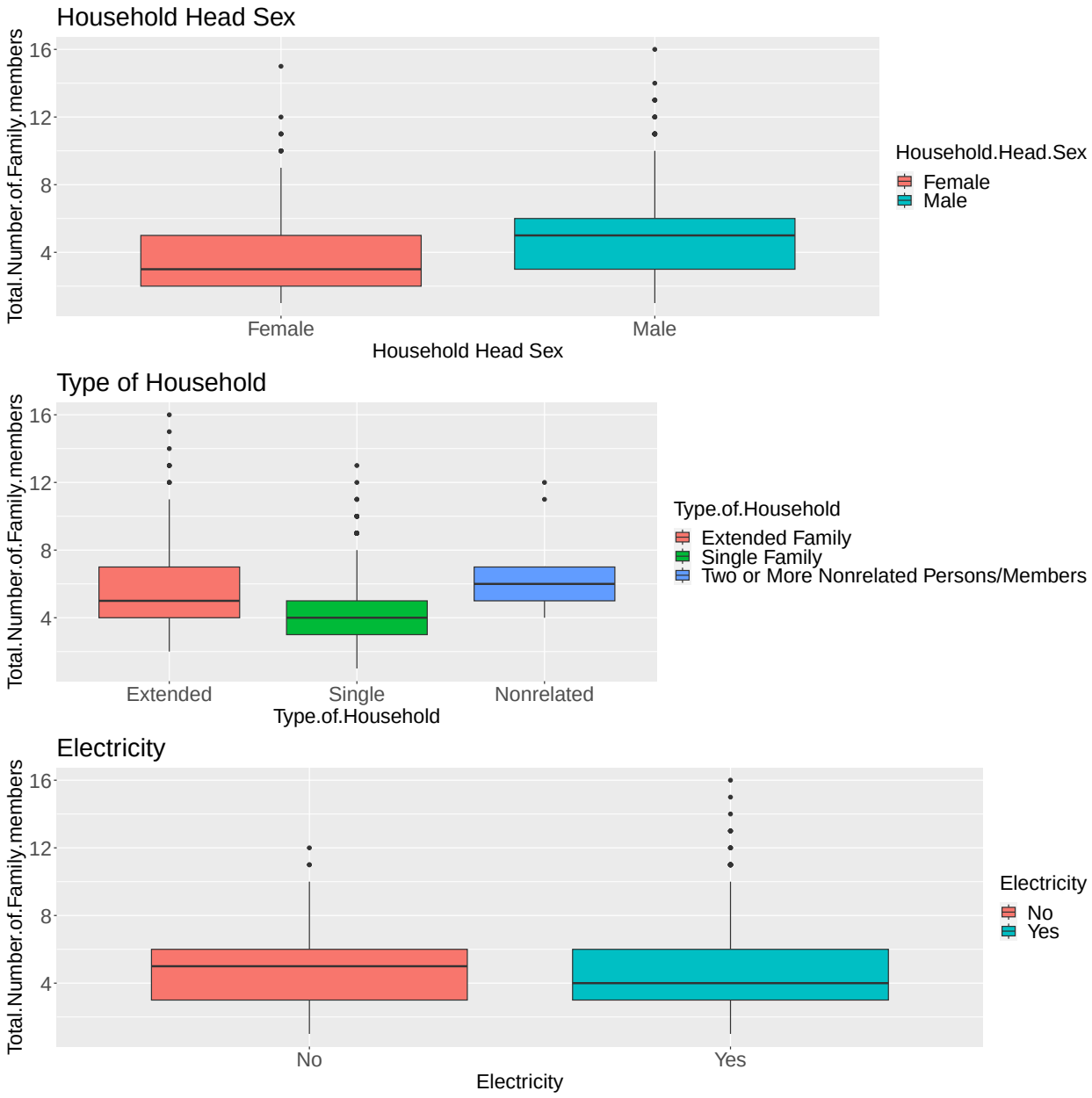

Figure 4: Boxplot

From the boxplots in Figure 4,

1. The median number of family members is higher in a household with a male head, than female head.
2. The median number of family members in a Two or More Nonrelated Persons/Members household than Extended Family household. Also, the minimum number of family members in a Two or More Nonrelated Persons/Members household is equivalent to the median value of a single member household.
3. The median value of the number of family members with no electricity is a bit higher than the number of members with electricity; though the spread seems to be the same.

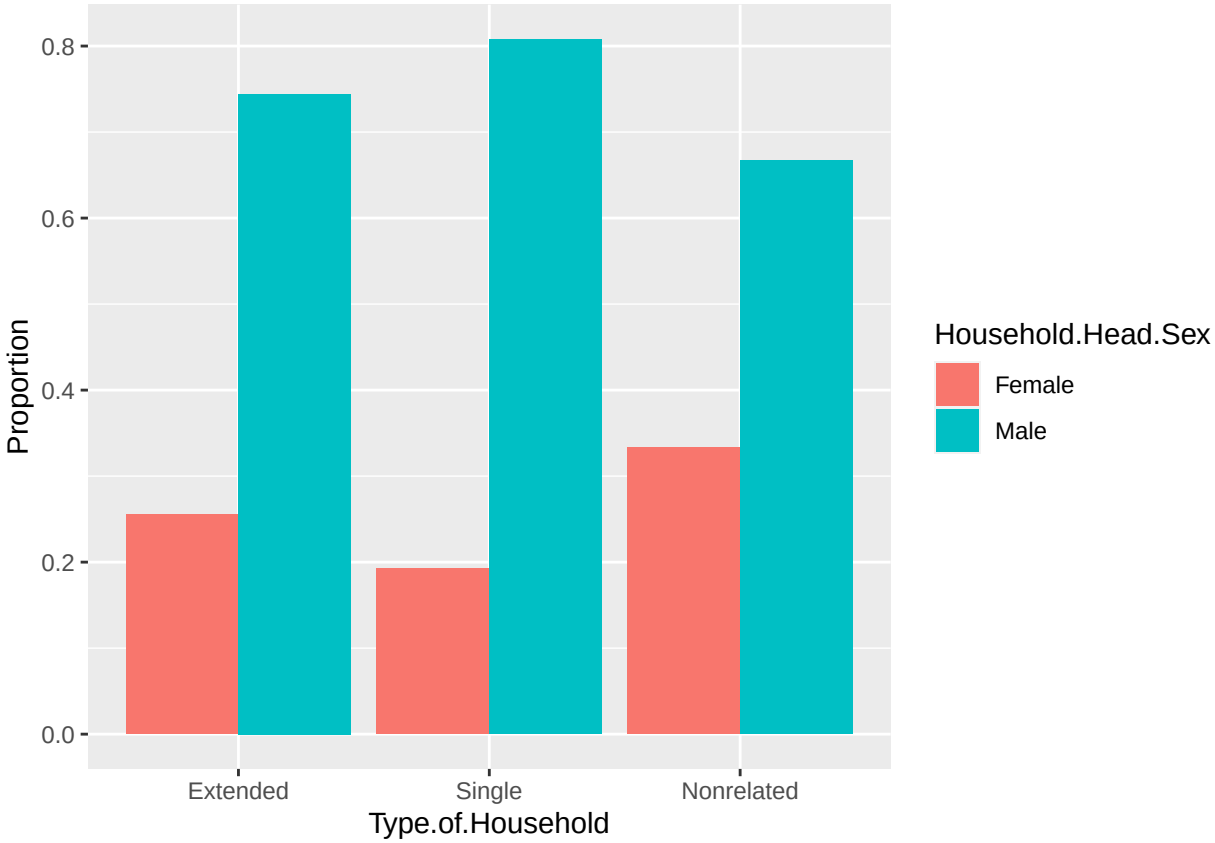

Figure 5: Barplot of Type of Household and Household Head Sex

Table 3: Proportion : Type of Household and Household Head Sex

Type.of.Household	Household.Head.Sex	n	Proportion
Extended Family	Female	145	25.57319
Extended Family	Male	422	74.42681
Single Family	Female	252	19.22197
Single Family	Male	1059	80.77803
Two or More Nonrelated Persons/Members	Female	3	33.33333
Two or More Nonrelated Persons/Members	Male	6	66.66667

From the Barplot in Figure 5, overall, the proportion is higher for male household heads than females.

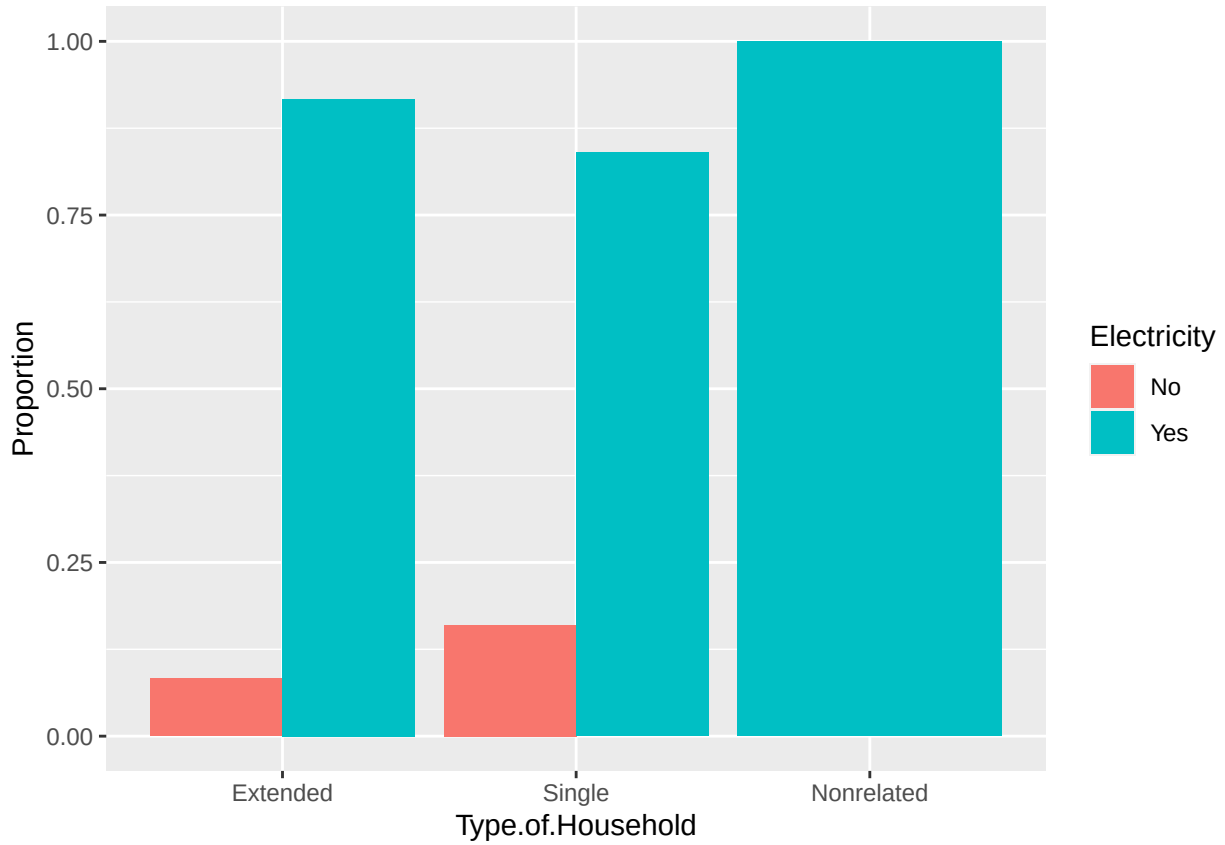

Figure 6: Barplot of Type of Household and Electricity

Table 4: Proportion : Type of Household and Electricity

Type.of.Household	Electricity	n	Proportion
Extended Family	No	47	8.289242
Extended Family	Yes	520	91.710758
Single Family	No	210	16.018307
Single Family	Yes	1101	83.981693
Two or More Nonrelated Persons/Members	Yes	9	100.000000

From the Barplot in Figure 6,

1. Households with Two or More Nonrelated Persons/Members always have Electricity.
2. Around 8% and 16% of the Extended family and Single family households have no electricity, respectively.

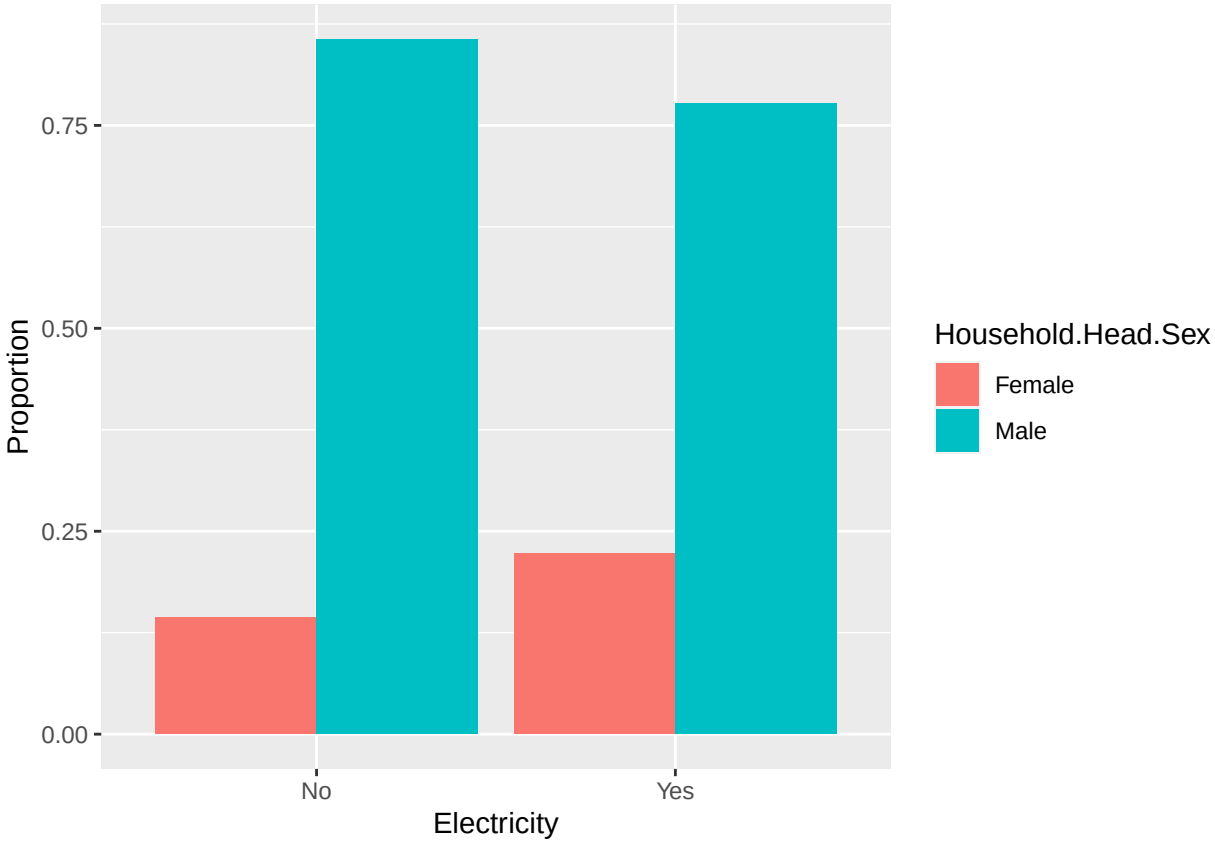

Figure 7: Barplot of Electricity and Househols Head Sex

Table 5: Proportion : Electricity and Household Head Sex

Electricity	Household.Head.Sex	n	Proportion
No	Female	37	14.39689
No	Male	220	85.60311
Yes	Female	363	22.26994
Yes	Male	1267	77.73006

From the Barplot in Figure 7,

1. Around 85% of households with no electricity have male heads.
2. Around 78% of households with electricity have male heads.

Formal Data Analysis

Logistic Regression

We will model the data using Logistic Regression, by converting Total.Number.of.Family.members to a binary variable. We create a new variable, Members.bi and set it to 0 if Total.Number.of.Family.members less than or equal to 4 and 1 otherwise.

Table 6: Logistic Model Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	17.2295791	921.5748744	0.0186958	0.9850838
Total.Household.Income	-0.0000025	0.0000015	-1.6848769	0.0920123
Total.Food.Expenditure	0.0000896	0.0000118	7.5772765	0.0000000
Household.Head.SexMale	0.9835387	0.2522537	3.8990055	0.0000966
Household.Head.Age	-0.0080901	0.0075547	-1.0708706	0.2842276
Type.of.HouseholdSingle Family	-18.3006885	921.5746674	-0.0198581	0.9841566
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-2.4936927	7590.5410399	-0.0003285	0.9997379
House.Floor.Area	-0.0050725	0.0033284	-1.5240002	0.1275087
House.Age	-0.0342706	0.0082897	-4.1341031	0.0000356
Number.of.bedrooms	0.0189622	0.1429920	0.1326103	0.8945016
ElectricityYes	0.5774051	0.3115998	1.8530344	0.0638774
Members.bi	17.5485400	761.6241869	0.0230409	0.9816176

Poisson Regression

For our initial analysis, Model 1, a main effects model was considered. Insignificant terms were dropped one by one based on p-value and arrived at Model 2. From Table 8, Model 2 did not seem to fit the data better. So, Model 1 was preferred.

Model 3, with all two-way interactions, was considered along with Model 4 and Model 4 which were updated from Model 3.

Table 7: Model 5 Coefficients

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.7300435	0.0521287	14.004624	0.0000000
Total.Household.Income	-0.0000003	0.0000001	-2.777951	0.0054703
Total.Food.Expenditure	0.0000120	0.0000007	17.107200	0.0000000
Household.Head.SexMale	0.4176668	0.0510811	8.176536	0.0000000
Total.Household.Income:Total.Food.Expenditure	0.0000000	0.0000000	-6.690897	0.0000000
Total.Food.Expenditure:Household.Head.SexMale	-0.0000033	0.0000006	-5.500775	0.0000000

Table 8: Compare Model 1 and 2

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1876	1398.35	NA	NA	NA
1878	1399.57	-2	-1.22	0.54

From, Tables 9,10 and 11, Model 5 seemed to fit the data best.

Table 9: Compare Model 1 and 3

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1876	1398.35	NA	NA	NA
1833	1231.15	43	167.2	0

Table 10: Compare Model 3 and 4

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1833	1231.15	NA	NA	NA
1874	1439.98	-41	-208.83	0

Table 11: Compare Model 4 and 5

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1874	1439.98	NA	NA	NA
1881	1554.18	-7	-114.2	0

Table 12: Confidence Interval

	2.5 %	97.5 %
(Intercept)	0.6271278	0.8314877
Total.Household.Income	-0.0000005	-0.0000001
Total.Food.Expenditure	0.0000107	0.0000134
Household.Head.SexMale	0.3179428	0.5182124
Total.Household.Income:Total.Food.Expenditure	0.0000000	0.0000000
Total.Food.Expenditure:Household.Head.SexMale	-0.0000044	-0.0000021

From Table 7 and 12, we can see that the p-value are less than 0.05 and the confidence intervals do not contain 0. Therefore, all the variables in Model 5 are significant.

Model 5 is given by:

$$
\begin{aligned}
&log(Total.Number.of.Family.members) = 0.7300435 \\
&+ -3.1495264 \times 10^{-7}\ Total.Household.Income \\
&+ 1.2043945 \times 10^{-5}\ Total.Food.Expenditure \\
&+ 0.4176668\ Household.Head.Sex \\
&+ -4.0753636 \times 10^{-12}\ Total.Household.Income : Total.Food.Expenditure \\
&+ -3.2602287 \times 10^{-6}\ Total.Food.Expenditure : Household.Head.Sex
\end{aligned}
$$

Conclusions

1. Significant variables are Income, Food Expenditure, and Household Head Sex.
2. Households with male heads tend to have more members but lower associated increases with Food Expenditure.
3. Households with higher incomes tend to have fewer members.
4. Higher Household Food expenditure tends to indicate more members