Group_1_Analysis_LB

Linbin Lai

1 Data Analysis

1.1 Load Package and Document

```
rm(list = ls())
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(MASS)
library(dplyr)
library(janitor)
library(tidyr)
FIES <- read.csv("dataset01.csv")
unique(FIES$Region)#ignore the region column
```

```
[1] "CAR"
```

```
FIES <- FIES %>%
  dplyr::select(Total.Number.of.Family.members, Total.Household.Income,
 ↪  Total.Food.Expenditure, Household.Head.Sex, Household.Head.Age,
 ↪  Type.of.Household, House.Floor.Area, House.Age, Number.of.bedrooms,
 ↪  Electricity)
FIES$Total.Number.of.Family.members <-
 ↪  as.factor(FIES$Total.Number.of.Family.members)
FIES$Household.Head.Sex <- as.factor(FIES$Household.Head.Sex)
FIES$Type.of.Household <- as.factor(FIES$Type.of.Household)
```

```
FIES$Electricity <- as.factor(FIES$Electricity)
levels(FIES$Electricity) <- c("No", "Yes")
FIES$Number.of.bedrooms <- as.factor(FIES$Number.of.bedrooms)
levels(FIES$Number.of.bedrooms) <- c("0", "1", "2", "3", "4", "5", "6",
 ↪ "7", "8", "9")
```

1.2 Analyze each element individually

1.2.1 `Total.Number.of.Family.members` vs `Total.Household.Income`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
 ↪ Total.Household.Income, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Annual
   ↪ household income (in Philippine peso)") +
  theme(legend.position = "none")
```

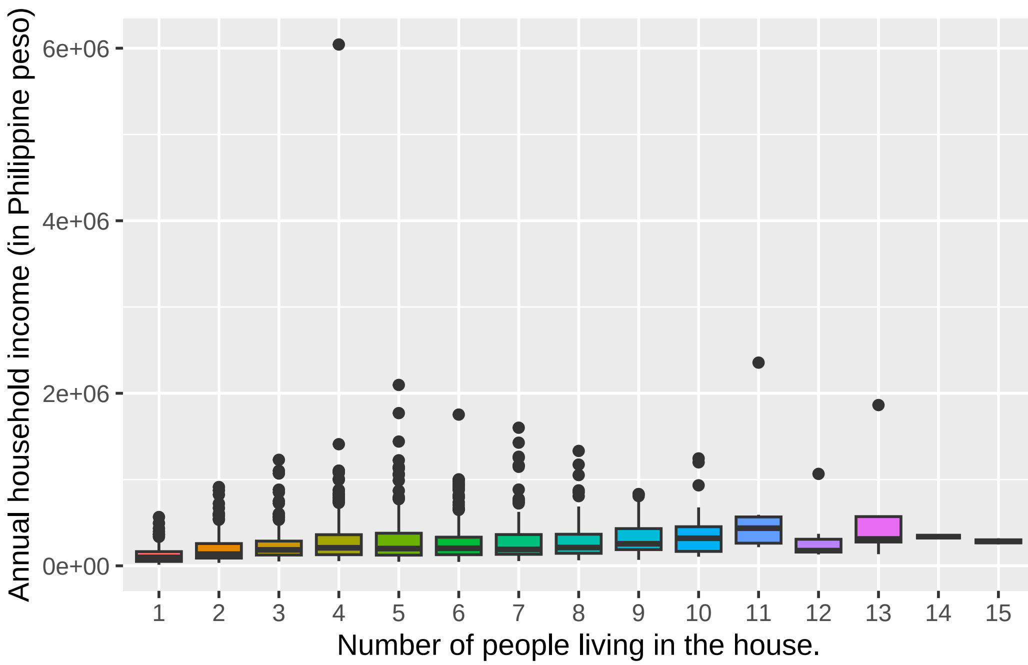

Figure 1: Annual household income by Number of family members.

From the boxplot, we can find that median annual income appears to be similar for most household sizes (number of people), particularly from one to six person households. Moreover, as household size increases, the number of households with unusually high incomes decreases.

1.2.2 `Total.Number.of.Family.members` vs `Total.Food.Expenditure`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
↪  Total.Household.Income, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Annual
   ↪  expenditure by the household on food") +
  theme(legend.position = "none")
```

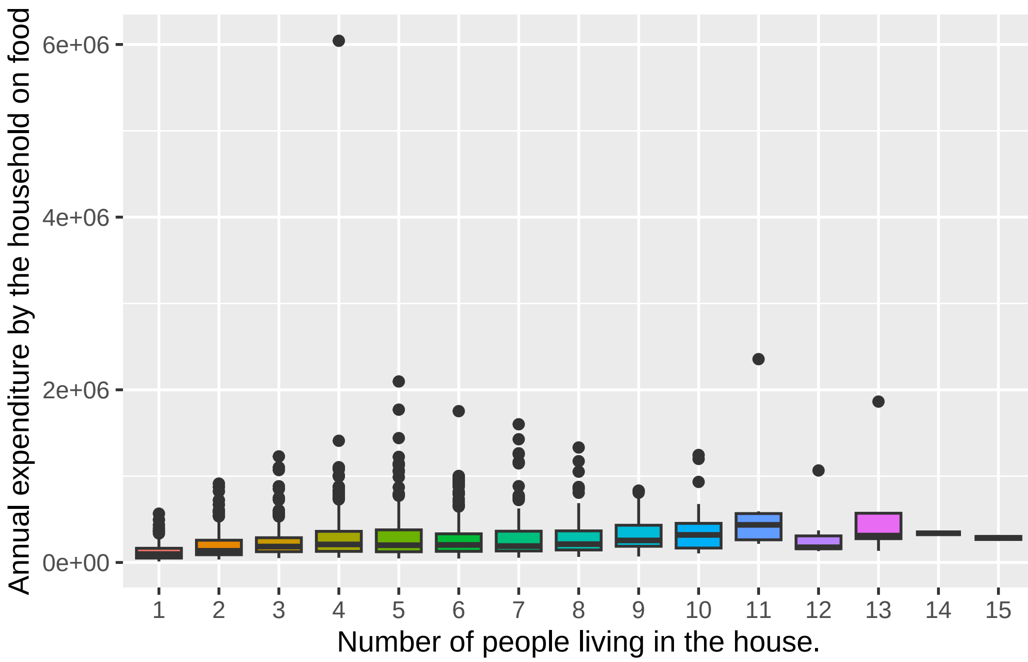

Figure 2: Annual expenditure by the household on food by Number of family members.

From this boxplot, median food expenditures appear to increase gradually as household size increases, especially from one to nine person households. There are fewer data points for larger households (e.g. more than 10 people), as can be seen from the width of the boxplot, with smaller box widths indicating smaller sample sizes for these household sizes.

1.2.3 `Total.Number.of.Family.members` vs `Household.Head.Sex`

```
FIES %>%
  tabyl(Household.Head.Sex, Total.Number.of.Family.members) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns() #To show original counts
```

```
Household.Head.Sex          1           2           3           4           5
            Female 14.6% (54) 13.6%  (50) 20.3%  (75) 17.3%  (64) 11.7%  (43)
              Male  5.5% (74)  9.8% (133) 13.3% (180) 17.8% (242) 18.5(251)
          6          7         8          9         10         11       12
 9.5%  (35) 5.7%  (21) 3.0% (11) 1.9%  (7) 0.5%  (2) 0.3%  (1) 0.8     (3)
14.2% (193) 8.8% (119) 5.6% (76) 2.9% (39) 1.8% (24) 1.0% (13) 0.5     (7)
      13       14       15
0.8% (3) 0.0% (0) 0.0                                                  (0)
0.1% (2) 0.1% (1) 0.1                                                  (2)
```

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, group =
↪  Household.Head.Sex)) +
  geom_bar(aes(y = ..prop.., fill = Household.Head.Sex), stat = "count",
  ↪  position = "dodge") +
  labs(x = "Number of people living in the house", y = "Proportion")
```

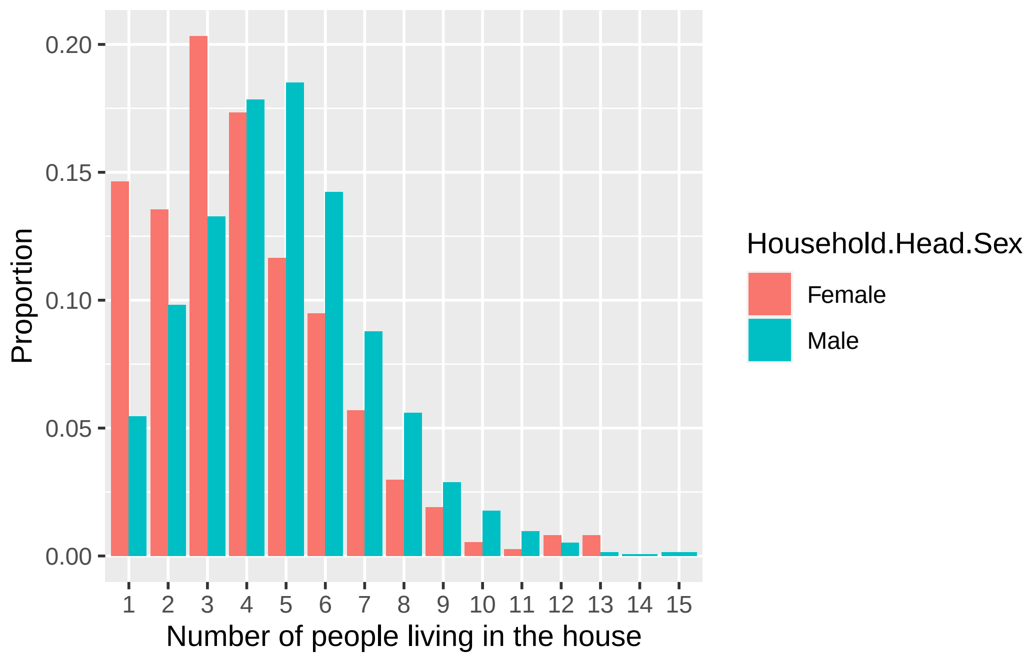

Figure 3: Barplot of Number of people living in the house by Head of the households sex.

For different family sizes, families headed by men generally account for a larger proportion, especially in families with three to seven people. Among one person, two person and three person households, the proportion of female-headed households is higher than the male-headed households.

1.2.4 `Total.Number.of.Family.members` vs `Household.Head.Age`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
 ↪  Household.Head.Age, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Head of the
   ↪  households age (in years)") +
  theme(legend.position = "none")
```

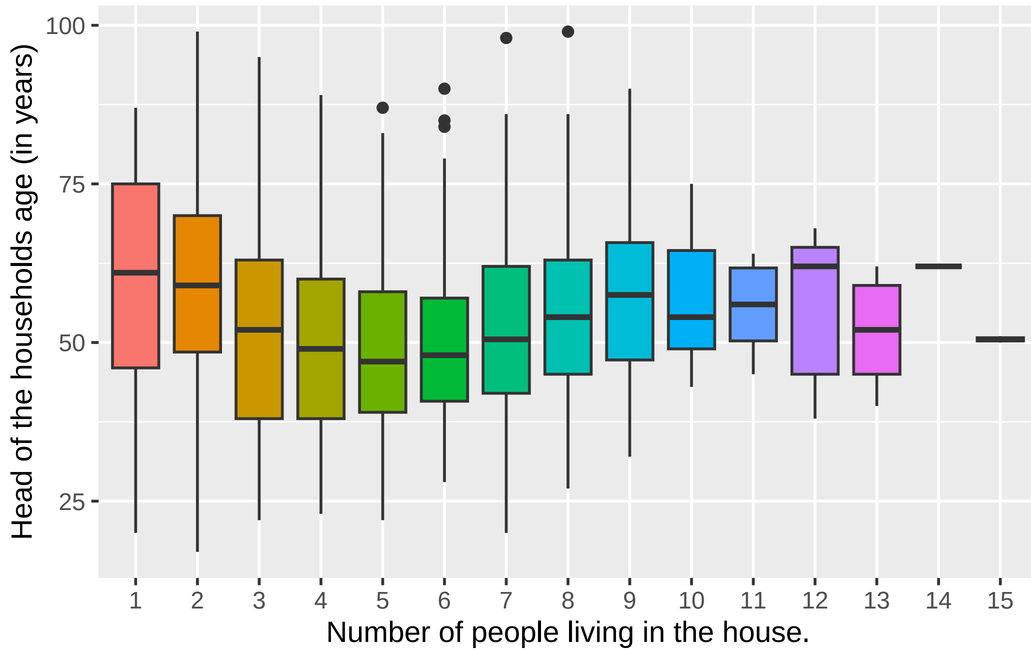

Figure 4: Head of the households age by Number of family members.

For different family sizes, the median age of the household head is relatively stable among different family sizes, mainly concentrated around 50 years old. So I think there is no obvious correlation trend between the age of the household head and the size of the family.

1.2.5 `Total.Number.of.Family.members` vs `Type.of.Household`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, group =
↪  Type.of.Household)) +
  geom_bar(aes(y = ..prop.., fill = Type.of.Household), stat = "count",
   ↪  position = "dodge") +
  labs(x = "Number of people living in the house", y = "Proportion")
```

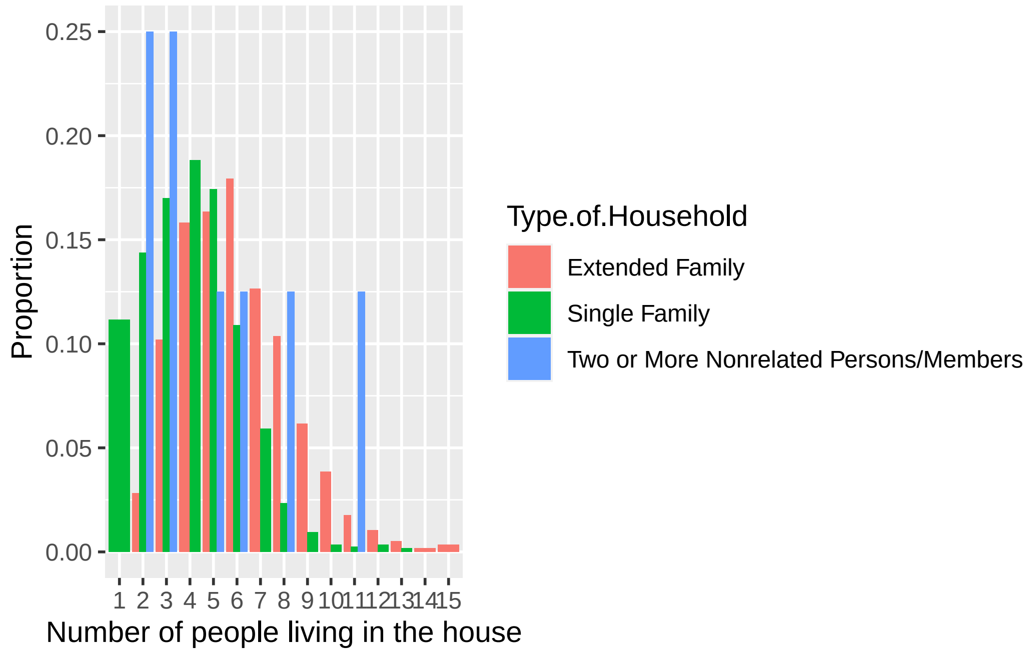

Figure 5: Barplot of Number of people living in the house by the relationship between the group of people living in the house.

There is a correlation between household type and the number of people living in the house, with extended households being more common in medium and large households, while single households predominate among small households. And for two or More Non-related Persons/Members, it has highly proportion in two person and three person household.

1.2.6 `Total.Number.of.Family.members` vs `House.Floor.Area`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
↪  House.Floor.Area, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Floor area of
  ↪  the house (in meter square)") +
  theme(legend.position = "none")
```

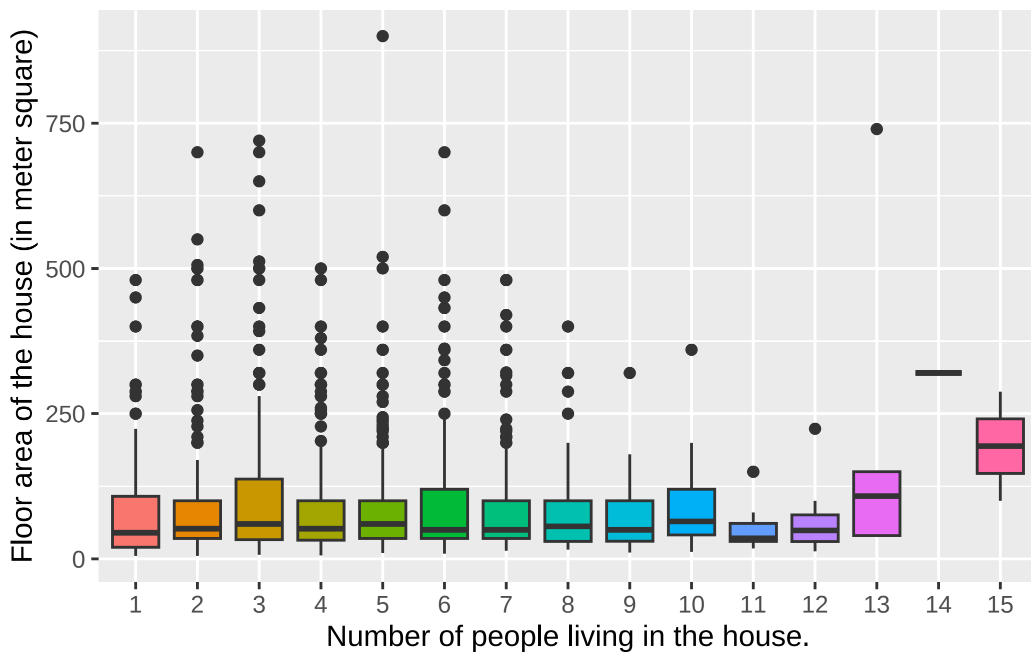

Figure 6: Floor area of the house by Number of family members.

Median house floor area appears to be relatively stable for most number of people living in the house, without increasing significantly as the number of residents increases. The graph shows a large number of outliers, indicating that some households have homes that are well outside the typical range for similar household sizes.Therefore, house floor area does not appear to increase significantly with number of people living in the house.

1.2.7 `Total.Number.of.Family.members` vs `House.Age`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, y =
↪  House.Age, fill = Total.Number.of.Family.members)) +
  geom_boxplot() +
  labs(x = "Number of people living in the house.", y = "Age of the
  ↪  building (in years)") +
  scale_fill_brewer(palette = "Set3") +#change to different color
  theme(legend.position = "none")
```

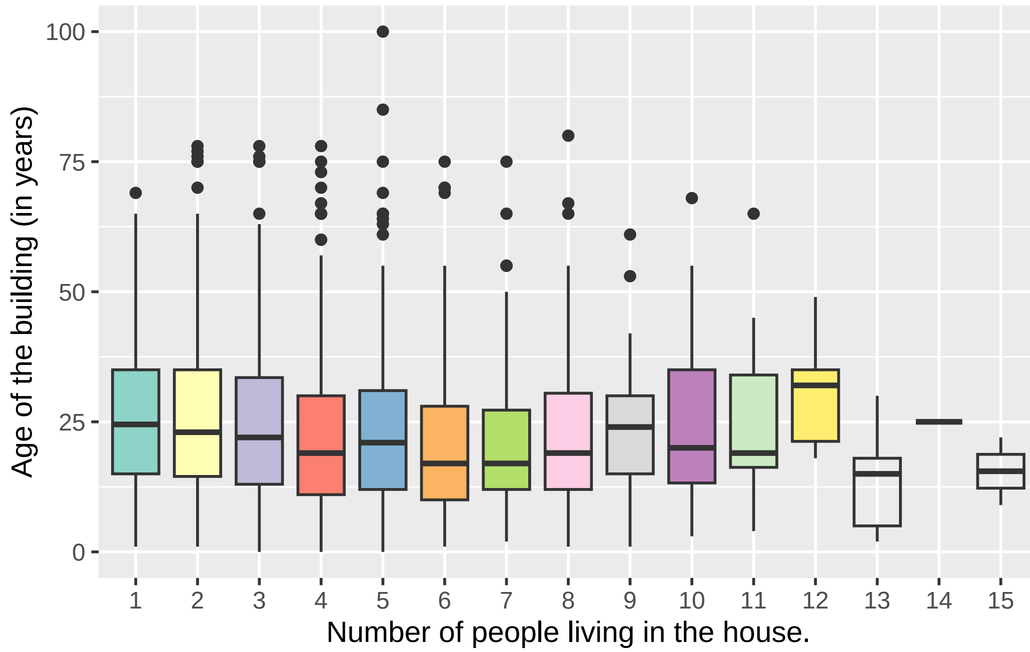

Figure 7: Age of the building by Number of family members.

The distribution range of house ages is relatively consistent, suggesting that the distribution of house ages is similar across household sizes. In conclude, there does not appear to be a direct relationship between house age and household size, with the median age remaining relatively stable across household sizes.

1.2.8 `Total.Number.of.Family.members` vs `Number.of.bedrooms`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, group =
↪  Number.of.bedrooms)) +
  geom_bar(aes(y = ..prop.., fill = Number.of.bedrooms), stat = "count",
  ↪  position = "dodge") +
  labs(x = "Number of people living in the house", y = "Proportion")
```

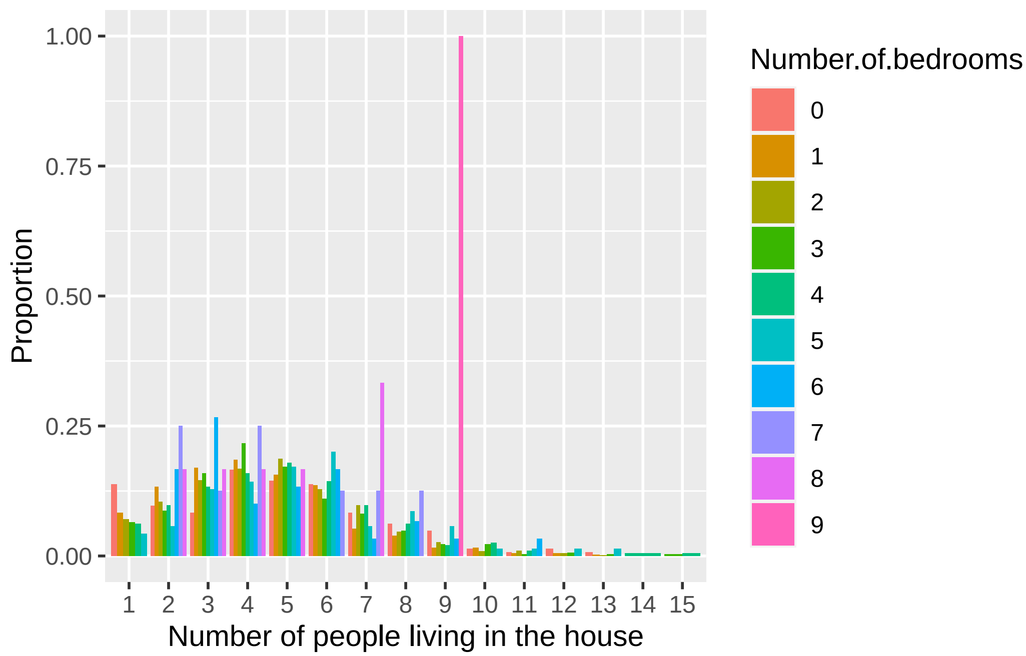

Figure 8: Barplot of Number of people living in the house by Number of bedrooms in the house.

As the number of residents increases, the median number of bedrooms gradually increases, but this growth is not linear.

1.2.9 `Total.Number.of.Family.members` vs `Electricity`

```
ggplot(data = FIES, aes(x = Total.Number.of.Family.members, group =
↪  Electricity)) +
  geom_bar(aes(y = ..prop.., fill = Electricity), stat = "count",
   ↪  position = "dodge") +
  scale_fill_brewer(palette = "Set1") +
  labs(x = "Number of people living in the house", y = "Proportion")
```

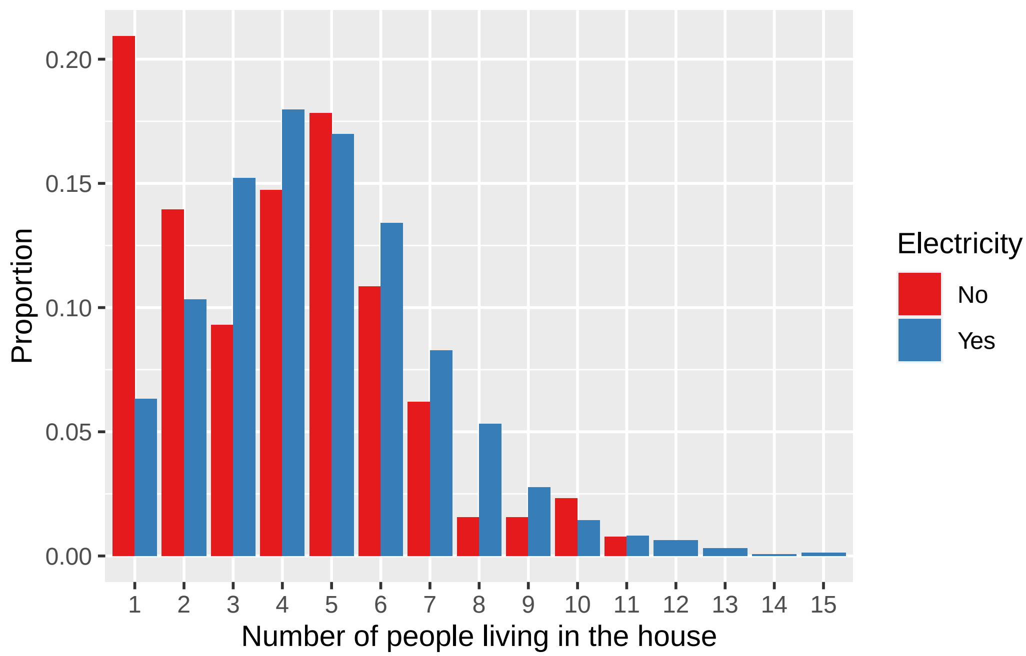

Figure 9: Barplot of Number of people living in the house by the electricity.

For most of the families, regardless of size, they have electricity. Among small scale households (especially single person households), the proportion of households without electricity is relatively high. As household size increases, the proportion of households without electricity gradually decreases. In larger households (seven people and above), almost all households have electricity supply.

1.3 Summary of Analyze each element individually

Overall, according to the separate analysis of explanatory variables, the variables that should be retained for now are:Annual household income, Annual expenditure by the household on food, Head of the households sex, Relationship between the group of people living in the house and Electricity.

1.4 Analysis of the overall part

```r
FIES <- FIES %>%
  dplyr::select(Total.Number.of.Family.members, Total.Household.Income,
↪   Total.Food.Expenditure, Household.Head.Sex, Type.of.Household,
↪   Number.of.bedrooms, Electricity)

FIES$Total.Number.of.Family.members <-
↪   as.numeric(as.character(FIES$Total.Number.of.Family.members))

model.FIES <- glm(Total.Number.of.Family.members ~
↪   Total.Household.Income + Total.Food.Expenditure + Household.Head.Sex
↪   + Type.of.Household + Electricity, data = FIES, family =
↪   poisson(link = "log"))

model.FIES %>%
  summary()
```

```

Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Type.of.Household +
    Electricity, family = poisson(link = "log"), data = FIES)

Coefficients:
                                                Estimate Std. Error
(Intercept)                                    1.185e+00  5.675e-02
Total.Household.Income                        -2.646e-07  5.541e-08
Total.Food.Expenditure                         5.017e-06  3.354e-07
Household.Head.SexMale                         2.480e-01  2.921e-02
Type.of.HouseholdSingle Family                -3.214e-01  2.390e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.265e-01  1.594e-01
ElectricityYes                                 4.664e-03  4.693e-02
                                              z value Pr(>|z|)
(Intercept)                                    20.885  < 2e-16 ***
Total.Household.Income                         -4.776 1.79e-06 ***
Total.Food.Expenditure                         14.958  < 2e-16 ***
Household.Head.SexMale                          8.489  < 2e-16 ***
Type.of.HouseholdSingle Family                -13.446  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.794    0.427
ElectricityYes                                  0.099    0.921
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1364.9  on 1718  degrees of freedom
AIC: 7041.4

Number of Fisher Scoring iterations: 4
```

`Type.of.HouseholdTwo or More Nonrelated Persons/Members`: The p value is 0.365302, indicating that the effect of this variable is not significant.
`Number.of.bedrooms (7 to 9 bedrooms)`: The p-values of these variables range from 0.129907 to 0.371571, indicating that their impact on the number of family members is not significant.
`Electricity Yes`: The p-value is 0.462128, which means that the presence or absence of electricity has no significant impact on the number of household members.

```
plot_model(model.FIES, show.values = TRUE,
           title = "", show.p = FALSE, value.offset = 0.25)
```

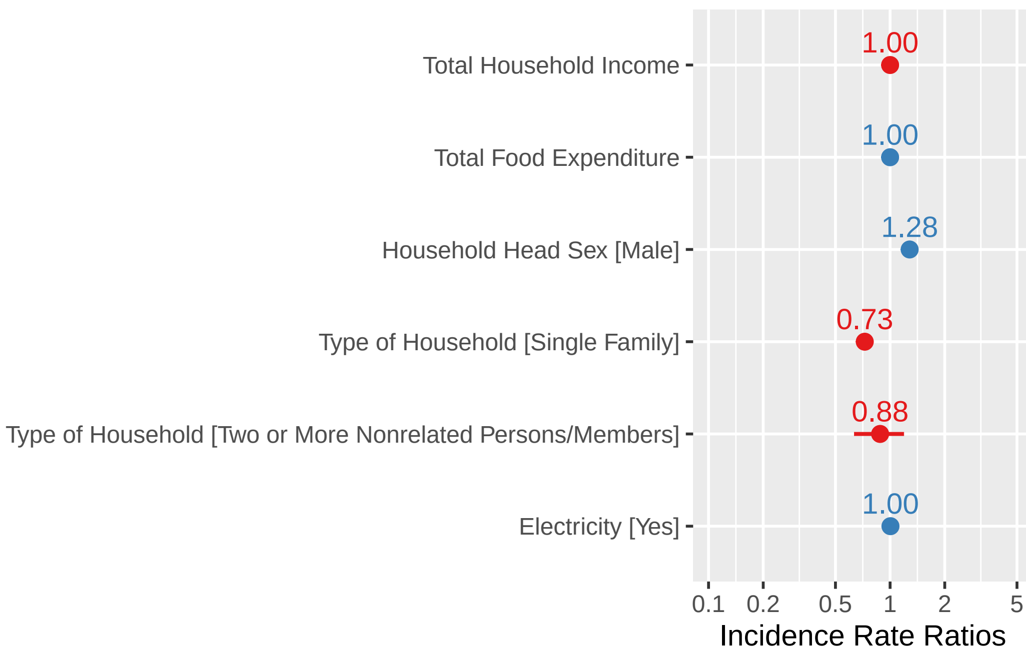

Figure 10: Rate ratios plot.

From total household income and total food expenditure, their rate ratios is exactly 1.00, they prove no effect on the number of family members. Household Head Sex's rate ratio is 1.28, which suggests that households headed by males have a 28% increase in the expected count of family members compared to households headed by females, assuming all other factors are held constant. Type of Household has the rate ratios about 0.72 indicates that single-family households have a 28% decrease in the expected count of family members compared to the reference household type (likely non-single family households). And for type of household With an rate ratios of 0.87, this suggests that households consisting of two or more non-related persons/members have a 13% decrease in the expected count of family members compared to the reference category. Number of Bedrooms suggesting a small decrease of 3% in the expected count of family members for each additional bedroom in the household.

1.5 Result

At last , we keep `Household.Head.Sex` and `Number.of.bedrooms` as explanatory variables:

$$Total.Number.of.Family.members = \beta_0 + \beta_1 \times \text{Household.Head.Sex}$$

```
FIES$Total.Number.of.Family.members <-
 ↪  as.numeric(as.character(FIES$Total.Number.of.Family.members))

model.FIES <- glm(Total.Number.of.Family.members ~ Household.Head.Sex,
 ↪  data = FIES, family = poisson(link = "log"))

model.FIES %>%
  summary()
```

```
Call:
glm(formula = Total.Number.of.Family.members ~ Household.Head.Sex,
    family = poisson(link = "log"), data = FIES)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.37813    0.02613  52.732  < 2e-16 ***
Household.Head.SexMale  0.20288    0.02889   7.022 2.19e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 2024.4  on 1724  degrees of freedom
Residual deviance: 1972.9  on 1723  degrees of freedom
AIC: 7639.5

Number of Fisher Scoring iterations: 4
```