DAS-Project2

Group04: Yiheng Yang, Yuanqing Zhang, Jie Li, Bing Gao, Ningxuan Zhang, Chao Z

1 Introduction

In order to find which household variables that influence the number of people living in a household. We use the datasets come from the Family Income and Expenditure Survey which is conducted in Philippines every three years. All the data were based on Soccsksargen, and the variables are shown in Table 1.

Table 1: Response variable and explanatory variables

	variable type	variable name	variable description
response variable	count	Total.Number.of.Family.members	Number of people living in the house
explanatory variables	categorical	Household.Head.Sex Type.of.Household Electricity	Head of the households sex Relationship between the group of people If the house have electricity
	numerical	Total.Household.Income Total.Food.Expenditure Household.Head.Age House.Floor.Area House.Age Number.of.bedrooms	Annual household income(in Philippine peso) Annual expenditure by the household on food Head of the household age(in years) Floor area of the house(in m^2) Age of the building(in years) Number of bedrooms in the house

2 Data Processing

2.1 Load the data

```
data=read.csv("dataset04.csv")
```

2.2 Get packages

```
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(tidyverse)
library(ggplot2)
library(MASS)
library(knitr)
library(tidyr)
library(gt)
library(janitor)
library(skimr)
library(kableExtra)
library(gridExtra)
```

2.3 Convert some categorical variables to factors

```
data$Household.Head.Sex=as.factor(data$Household.Head.Sex)
data$Type.of.Household=as.factor(data$Type.of.Household)
data$Electricity=as.factor(data$Electricity)
levels(data$Electricity)=c("No","Yes")
data$Number.of.bedrooms=as.factor(data$Number.of.bedrooms)
levels(data$Number.of.bedrooms)=c("0","1","2","3","4","5","6","7")
```

3 Exploratory Data Analysis

3.1 Summary of response variable

```
# Create a table to summarize the characteristics of the response variables
data%>%summarize('Mean' = mean(Total.Number.of.Family.members),
'Median' = median(Total.Number.of.Family.members),
'St.Dev' = sd(Total.Number.of.Family.members),
'Variance'=var(Total.Number.of.Family.members),
'Min' = min(Total.Number.of.Family.members),
'Max' = max(Total.Number.of.Family.members),
'IQR' = quantile(Total.Number.of.Family.members,0.75)
-quantile(Total.Number.of.Family.members,0.25),
'Sample_size' = n())%>%
  gt()%>%
  fmt_number(decimals=2)%>%
  cols_label(
Mean = html("Mean"),
Median = html("Median"),
St.Dev = html("Std. Dev"),
Variance=html("Variance"),
Min = html("Minimum"),
Max = html("Maximum"),
IQR = html("Interquartile Range"),
Sample_size = html("Sample Size"))
```

Mean	Median	Std. Dev	Variance	Minimum	Maximum	Interquartile Range	Sample Size
4.53	4.00	2.22	4.91	1.00	19.00	3.00	2, 122.00

We can see from this numerical summary, the mean of number of family members is 4.53 and the variance is 4.91. If variance is bigger than mean, we can determine that we have overdispersion. We will investigate this phenomenon later.

3.2 Summary of categorical explanatory variables

```
# Select the categorical explanatory variables
data_categorical=data%>%
  dplyr::select("Household.Head.Sex","Type.of.Household","Electricity")
```

```r
# Create a table to summarize the characteristics of the categorical explanatory variables
summary_table_categorical <-summary(data_categorical)
summary_table_categorical[is.na(summary_table_categorical)] <- ""
kable(summary_table_categorical,na.strings = "")
```

Household.Head.Sex	Type.of.Household	Electricity
Female: 362	Extended Family : 585	No : 363
Male :1760	Single Family :1531	Yes:1759
	Two or More Nonrelated Persons/Members: 6	

The numerical summary shows that male owners, single families and households with electricity account for a major proportion.

3.3 Summary of numerical explanatory variables

```r
# Create a table to summarize the characteristics of the numerical explanatory variables
data_numerical=data[,c(1,3,5,7,8,9,10)]
data_numerical$Number.of.bedrooms=as.numeric(as.character(data_numerical$Number.of.bedroom
my_skim <- skim_with(numeric = sfl(hist = NULL),
                    base = sfl(n = length))
my_skim(data_numerical) %>%
  transmute(Variable=skim_variable, Sample_size = n, Mean=numeric.mean, St.Dev=numeric.sd,
            Min=numeric.p0, Median=numeric.p50,  Max=numeric.p100,
            IQR = numeric.p75-numeric.p50) %>%
  kable(format.args = list(big.mark = ","), digits=2) %>%
  kable_styling(font_size = 10, latex_options = "hold_position")
```

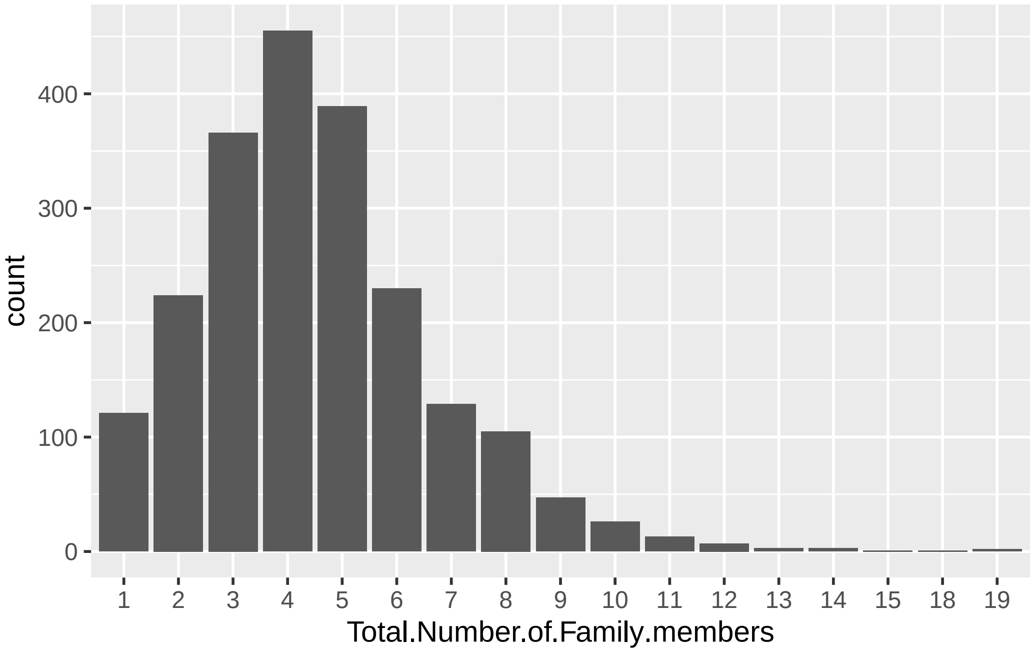
Variable	Sample_size	Mean	St.Dev	Min	Median	Max	IQR
Total.Household.Income	2,122	182,984.80	228,231.07	15,204	120,362.0	3,168,662	74,314.00
Total.Food.Expenditure	2,122	71,738.09	44,938.17	7,783	63,305.5	729,606	24,496.75
Household.Head.Age	2,122	49.28	14.16	9	48.0	99	11.00
Total.Number.of.Family.members	2,122	4.53	2.22	1	4.0	19	2.00
House.Floor.Area	2,122	35.74	34.67	5	26.5	450	13.50
House.Age	2,122	16.30	11.09	0	14.0	75	7.00
Number.of.bedrooms	2,122	1.77	1.00	0	2.0	7	0.00

3.4 Graphical summaries

As we want to plot a histogram with x axis to be number of family members, so we need to change this variable to be a factor.

```
# Convert the column "Total.Number.of.Family.members" to factor type
data$Total.Number.of.Family.members=as.factor(data$Total.Number.of.Family.members)
```

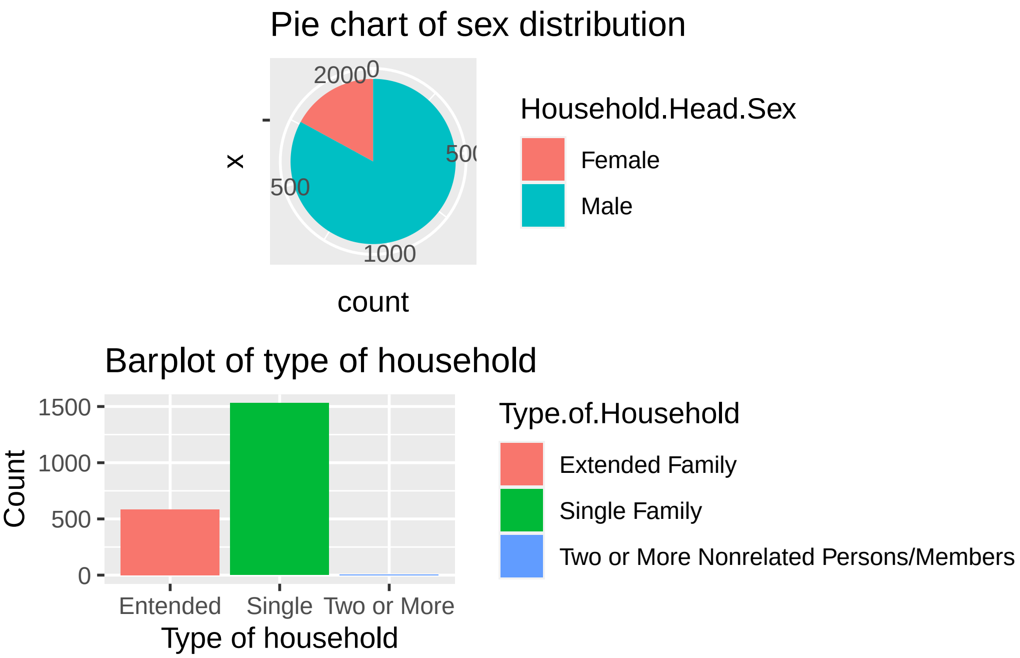
```
# Plot a histogram to show the distribution of response variable
#| label: fig-histogram_response
#| fig-cap: histogram of response variable
#| fig-align: center
#| message: false
ggplot(data=data,aes(x=Total.Number.of.Family.members))+geom_bar()
```



The **?@fig-histogram_response** shows that household with four family members accounts for the largest proportion. Most of the data is consisted of families with three to five family members.

```
p1=ggplot(data=data,aes(y=Total.Household.Income))+geom_boxplot()+labs(y="Total household
p2=ggplot(data=data,(aes(y=Total.Food.Expenditure)))+geom_boxplot()+labs(y="Total food exp
p3=ggplot(data=data,aes(x="",fill=Household.Head.Sex))+geom_bar(width=1)+coord_polar(theta
p4=ggplot(data=data,aes(y=Household.Head.Age))+geom_boxplot()+labs(y="Household head age",
p5=ggplot(data=data,aes(x=Type.of.Household))+geom_bar(aes(fill=Type.of.Household))+scale_
p6=ggplot(data=data,aes(y=House.Floor.Area))+geom_boxplot()+labs(y="House floor area",titl
p7=ggplot(data=data,aes(y=House.Age))+geom_boxplot()+labs(y="House age",title="Boxplot of 
p8=ggplot(data=data,aes(x=Number.of.bedrooms))+geom_bar(aes(fill=Number.of.bedrooms))+labs
p9=ggplot(data=data,aes(x=Electricity))+geom_bar(aes(fill=Electricity))+labs(y="Count",tit
```

```
# Arrange the plots in a grid layout for display
#| label: fig-piechart_sex
#| fig-cap: Pie chart of Household.Head.Sex
#| fig-align: center
#| message: false
grid.arrange(p3,p5,ncol=1)
```



```
grid.arrange(p8,p9,ncol=2)
```

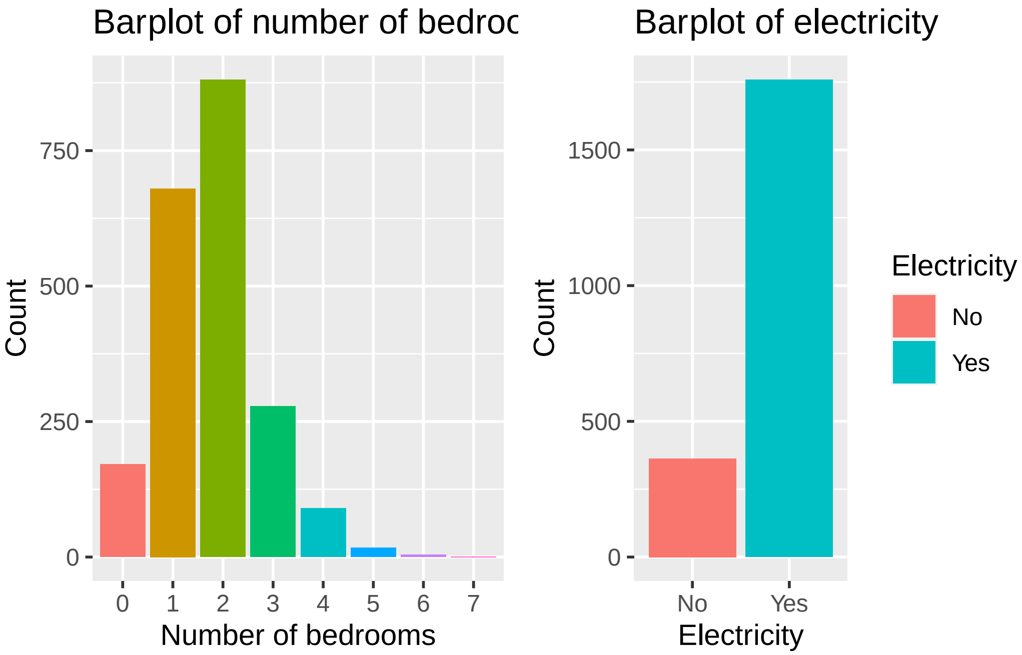


Figure 1: Barplots of some explanatory variables

```
grid.arrange(p1,p2,p4,p6,p7,ncol=3)
```

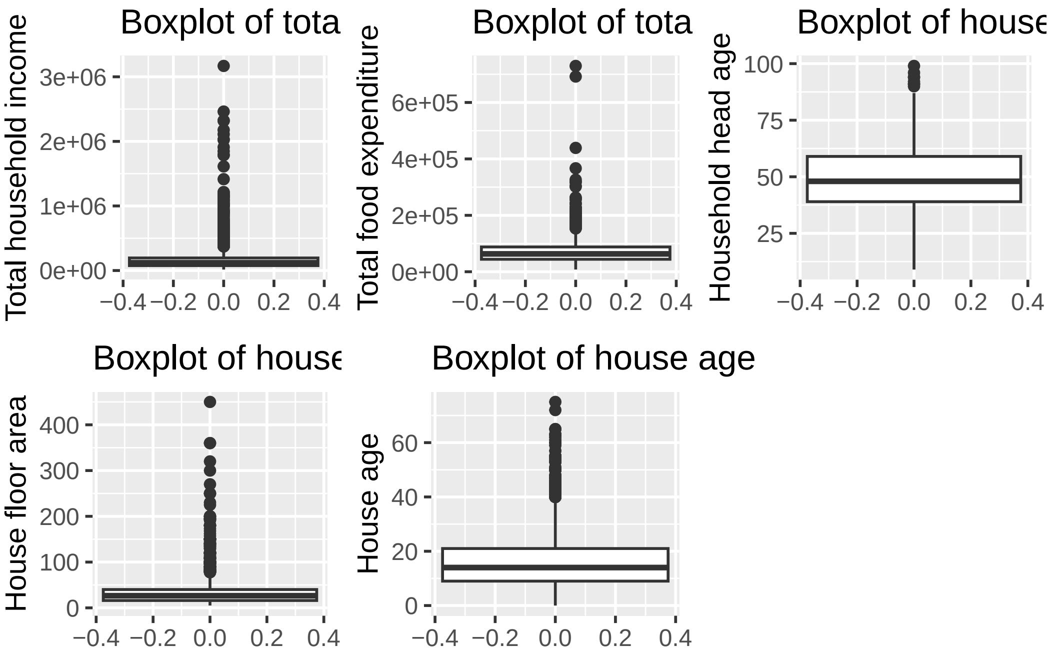


Figure 2: Boxplot of some explanatory variables

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=Total.Household.Income,
                     fill=Total.Number.of.Family.members))+geom_boxplot()+
                     theme(legend.position = "none")+
                     labs(x="Number of Family Members",
                                       y="Total Household Income")
```

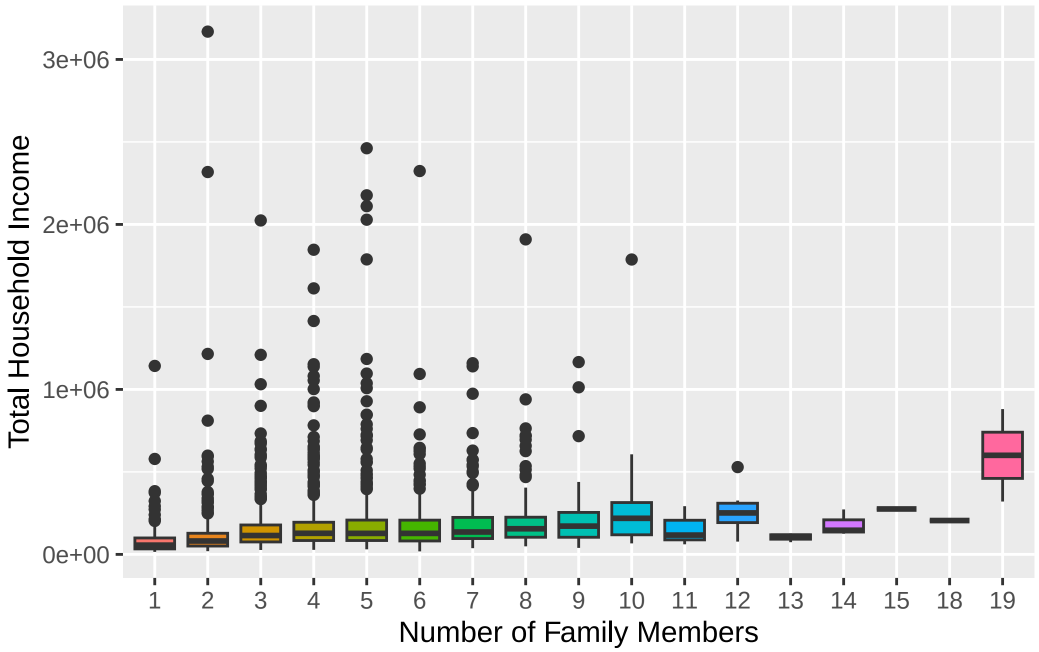


Figure 3: Income of families with different number of family members

We can see from the Figure 3 that the median of household income increase as number of family members increase.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=Total.Food.Expenditure,
                     fill=Total.Number.of.Family.members))+geom_boxplot()+
                     theme(legend.position = "none")+
                  labs(x="Number of Family Members",y="Total food expenditure")
```

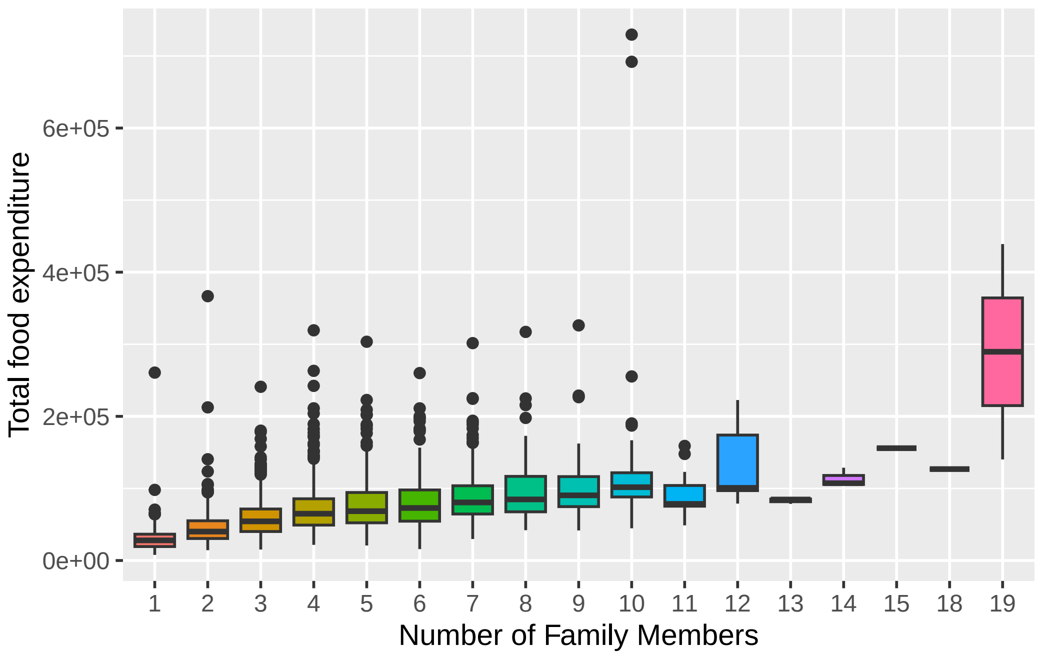


Figure 4: Food expenditure of families with different number of family members

The Figure 4 indicates that median increase significantly as the number of family members increase. Household with 19 members have the largest variance in food expenditure.

```
frequency_sex <- data%>%
  tabyl(Household.Head.Sex,Total.Number.of.Family.members)%>%
  adorn_percentages()%>%
  adorn_pct_formatting()%>%
  adorn_ns()
kable(frequency_sex)
```

Household.Head.Sex	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	18	19
Female	12.4% (45)	19.1% (69)	25.4% (92)	16.0% (58)	11.0% (40)	7.7% (28)	2.5% (9)	3.3% (12)	1.7% (6)	0.3% (1)	0.0% (0)	0.0% (0)	0.3% (1)	0.0% (0)	0.0% (0)	0.0% (0)	0.3% (1)
Male	4.3% (76)	8.8% (155)	15.6% (274)	22.6% (397)	19.8% (349)	11.5% (202)	6.8% (120)	5.3% (93)	2.3% (41)	1.4% (25)	0.7% (13)	0.4% (7)	0.1% (2)	0.2% (3)	0.1% (1)	0.1% (1)	0.1% (1)

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                     group=Household.Head.Sex))+geom_bar(aes(y=..prop..,
                    fill=Household.Head.Sex),position="dodge")+
                    labs(x="Number of Family Members",y="Proportion")
```

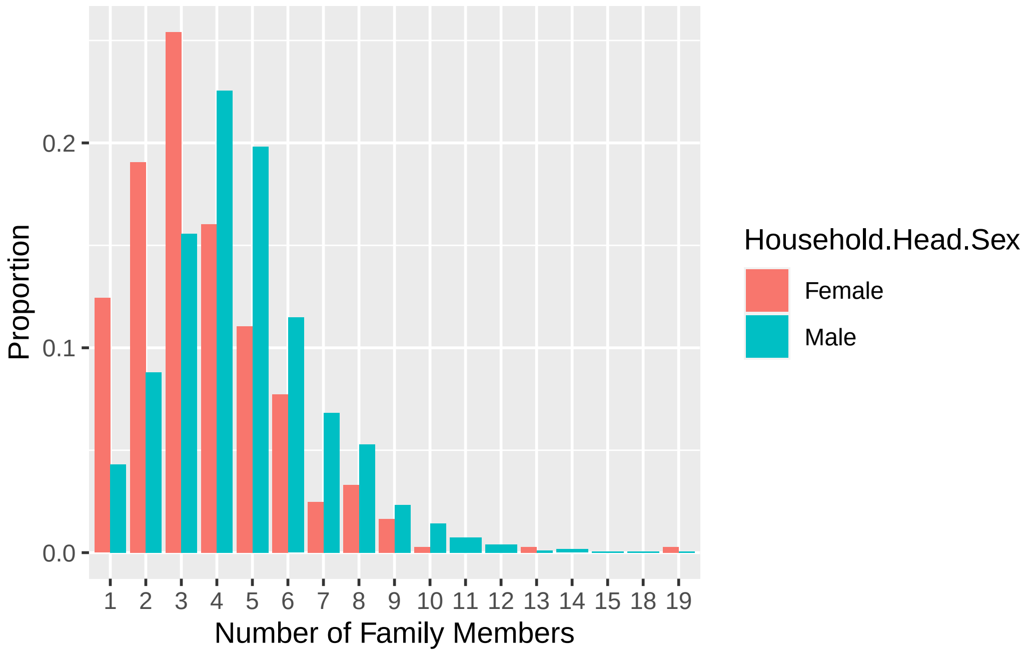


Figure 5: Head sex proportion for different size of households

We can see from the Figure 5, for those small sized households, the proportion is much higher for females than for males. However, this situation does not exist for those household with four or more family members.

```
ggplot(data=data,aes(x=Household.Head.Sex,
                     y=as.numeric(as.character(Total.Number.of.Family.members))))+
  geom_boxplot(aes(fill=Household.Head.Sex))+labs(x="Household head sex",
                                                   y="Number of family members")
```

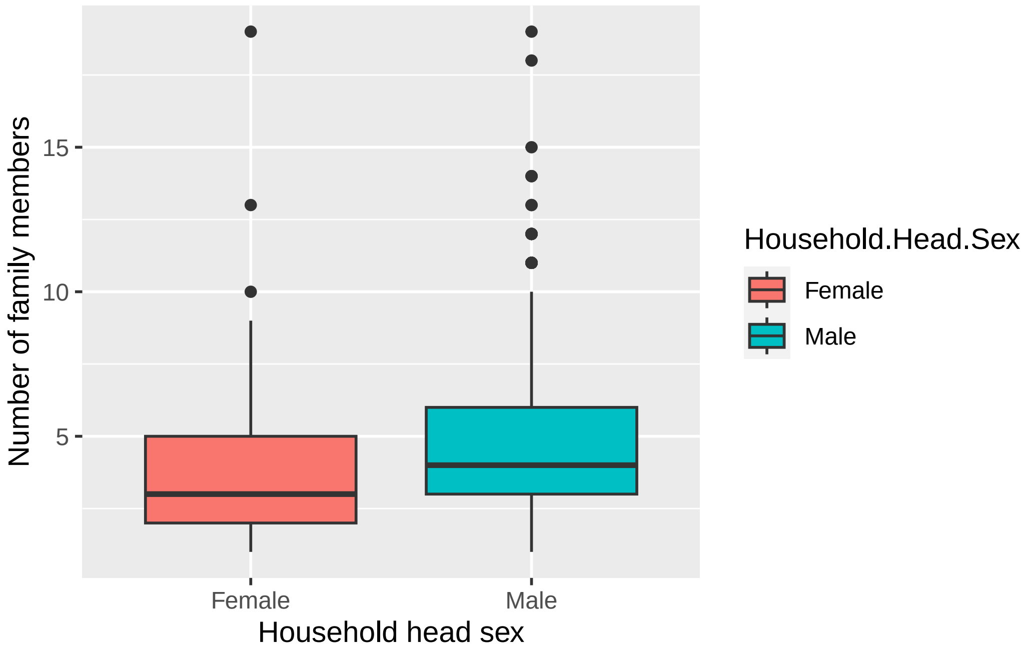


Figure 6: Number of family members by sex

We can conclude from the Figure 6 that households tend to have more family members if their owner is male.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=Household.Head.Age,fill=Total.Numb
```

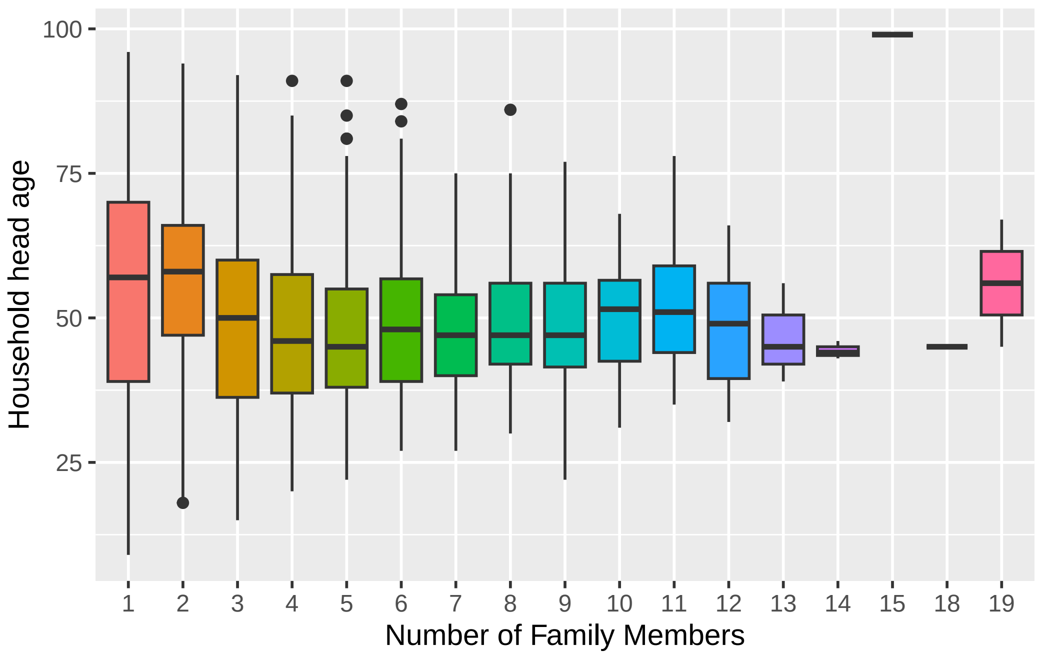


Figure 7: Head age for different size households

As shown in Figure 7, for different size of households, the median of household head age remain at a constant level around 50.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                     group=Type.of.Household))+
  geom_bar(aes(y=..prop..,fill=Type.of.Household))+
  labs(x="Number of Family Members",y="Proportion")
```

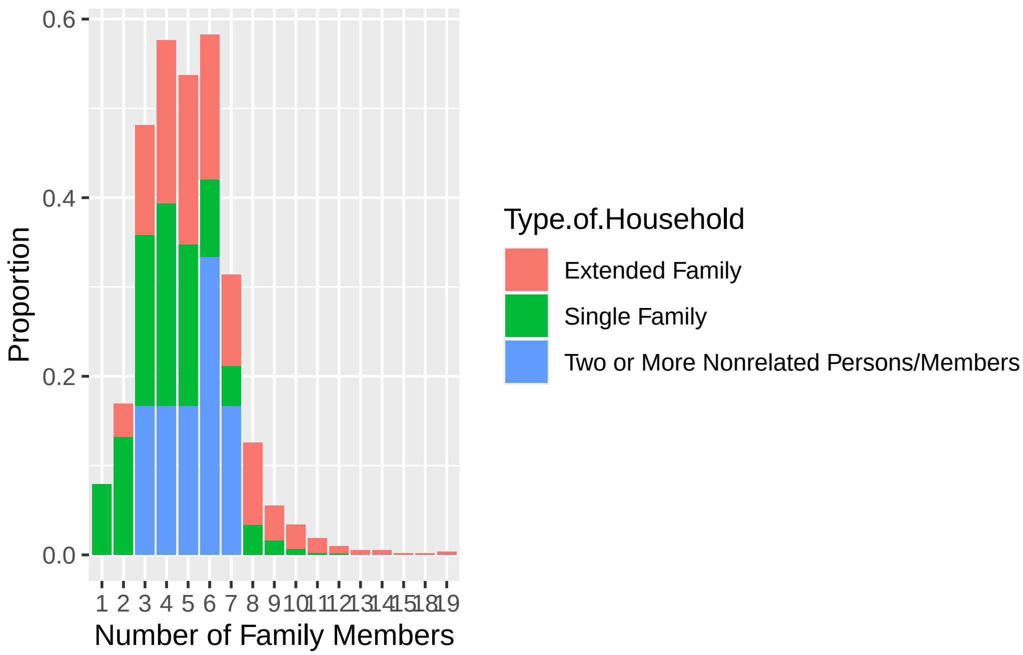


Figure 8: Type of household in different size of households

From we Figure 8 can see that these families with two or more nonrelated members only exist in medium size household. As total family members increase more than 8, single family account for a very small proportion.

```
ggplot(data=data,aes(x=Type.of.Household,
                     y=as.numeric(as.character(Total.Number.of.Family.members))))+
  geom_boxplot(aes(fill=Type.of.Household))+
  scale_x_discrete(labels=c("Extended","Single","Two or more"))+
  labs(x="Type of household",y="Number of family members")+
  theme(legend.position = "bottom")
```

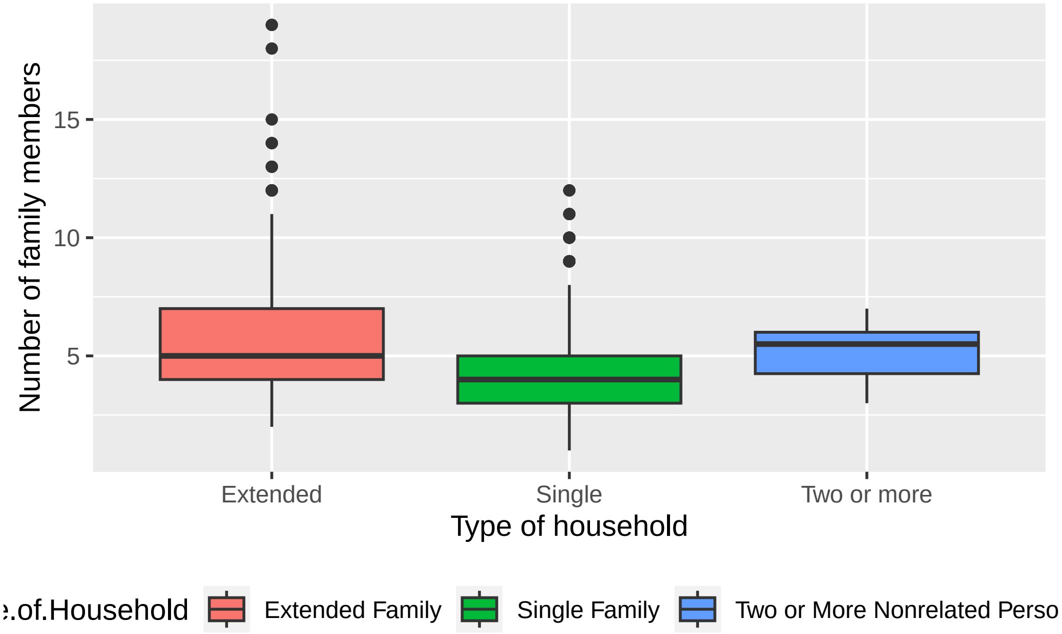


Figure 9: Number of family members by type of household

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                     y=House.Floor.Area,fill=Total.Number.of.Family.members))+
  geom_boxplot()+theme(legend.position = "none")+
  labs(x="Number of Family Members",y="House floor area")
```

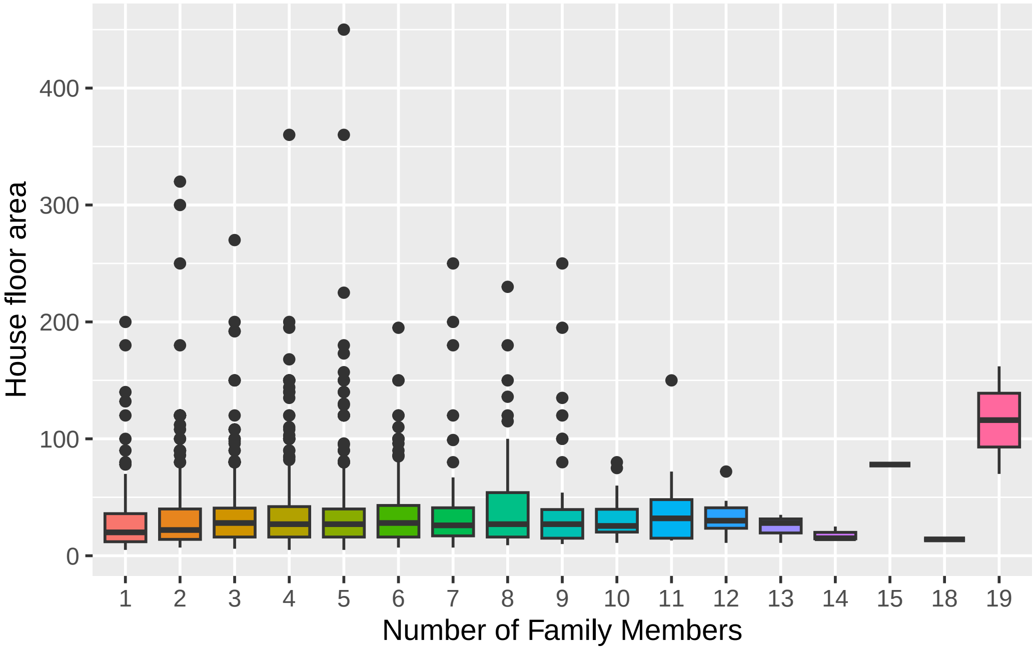


Figure 10: House floor area for different size of households

As shown in Figure 10, there are a few outliers for different sizes of households, . And the median of house floor area seems to be stable as number of family members increase.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,y=House.Age,
                     fill=Total.Number.of.Family.members))+
  geom_boxplot()+theme(legend.position = "none")+
  labs(x="Number of Family Members",y="House age")
```

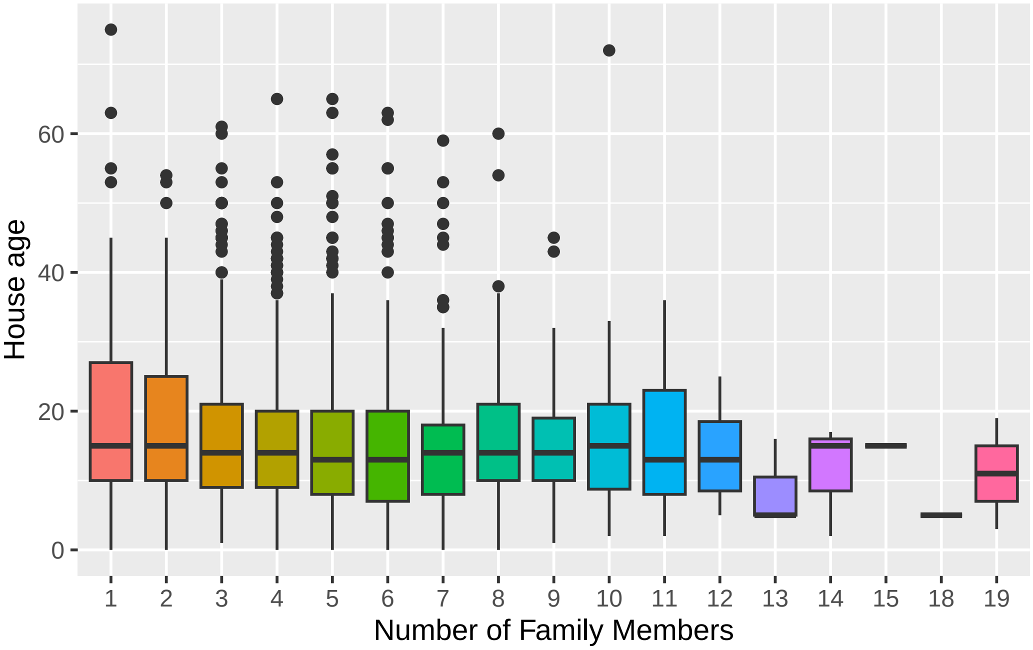


Figure 11: House age for different sizes of households

The median house age of different sizes of households are less than 20 years, which is relatively stable as number of family members increase. (Figure 11)

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                            group=Number.of.bedrooms))+
  geom_bar(aes(y=..prop..,fill=Number.of.bedrooms))+
  labs(x="Number of Family Members",y="Proportion")
```

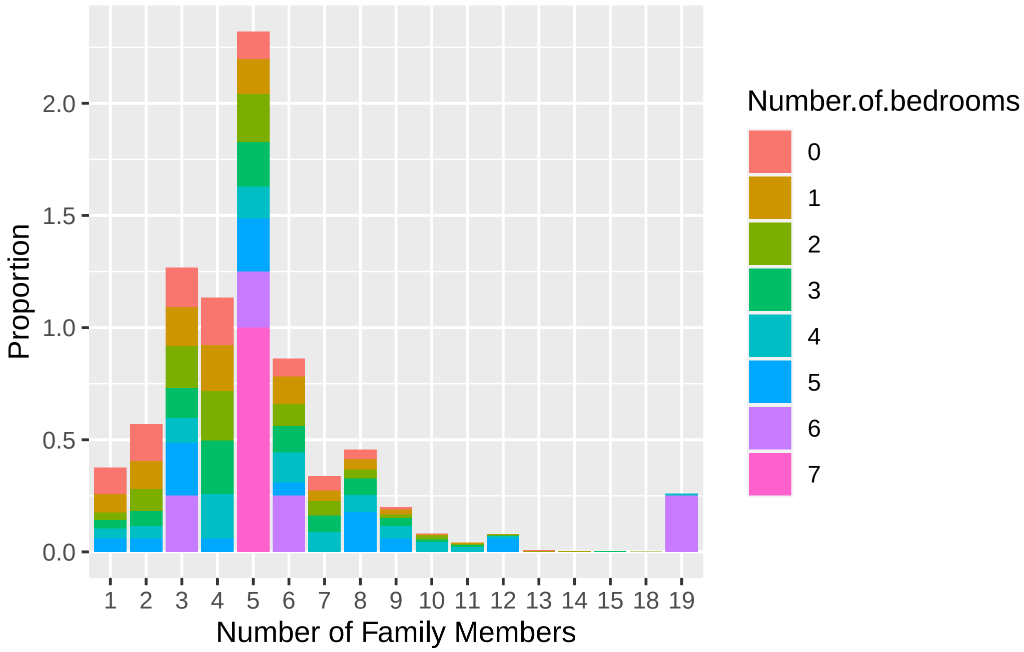


Figure 12: Number of bedrooms by different sizes of households

As the number of family members increases, number of bedrooms increase, but for household with 5 family members, proportion of 7 bedrooms is incredibly high.

```
ggplot(data=data,aes(x=Total.Number.of.Family.members,
                     group=Electricity))+
  geom_bar(aes(y=..prop..,fill=Electricity),position="dodge")+
  labs(x="Number of Family Members",y="Proportion")
```

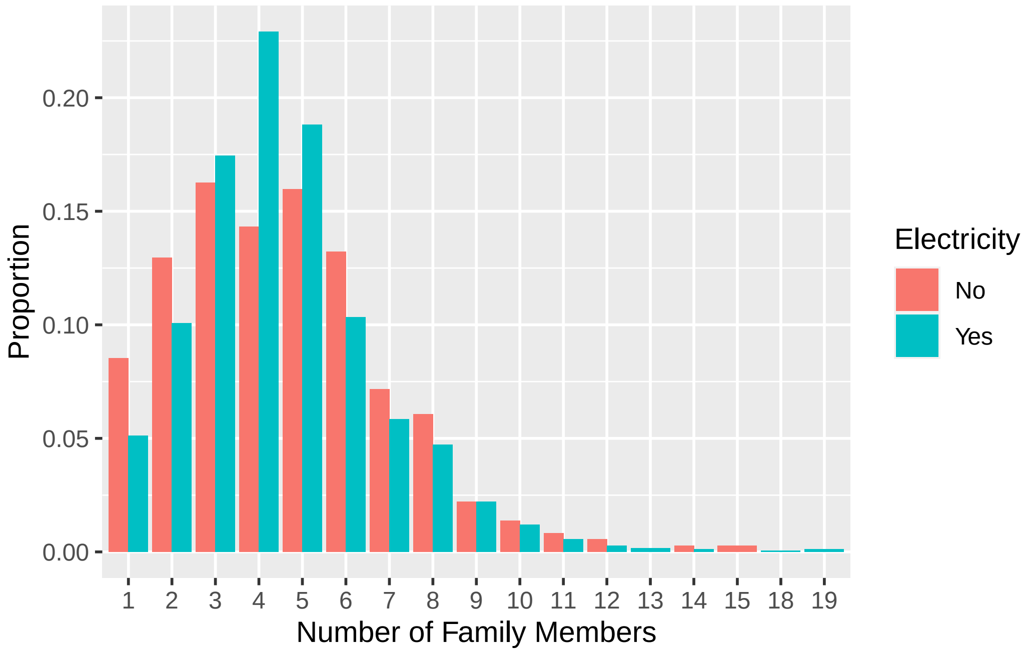


Figure 13: Electricity by different sizes of households

For those small size households, the proportion without electricity is relatively high.

```
ggplot(data=data,aes(x=Electricity,y=as.numeric(as.character(Total.Number.of.Family.member
```

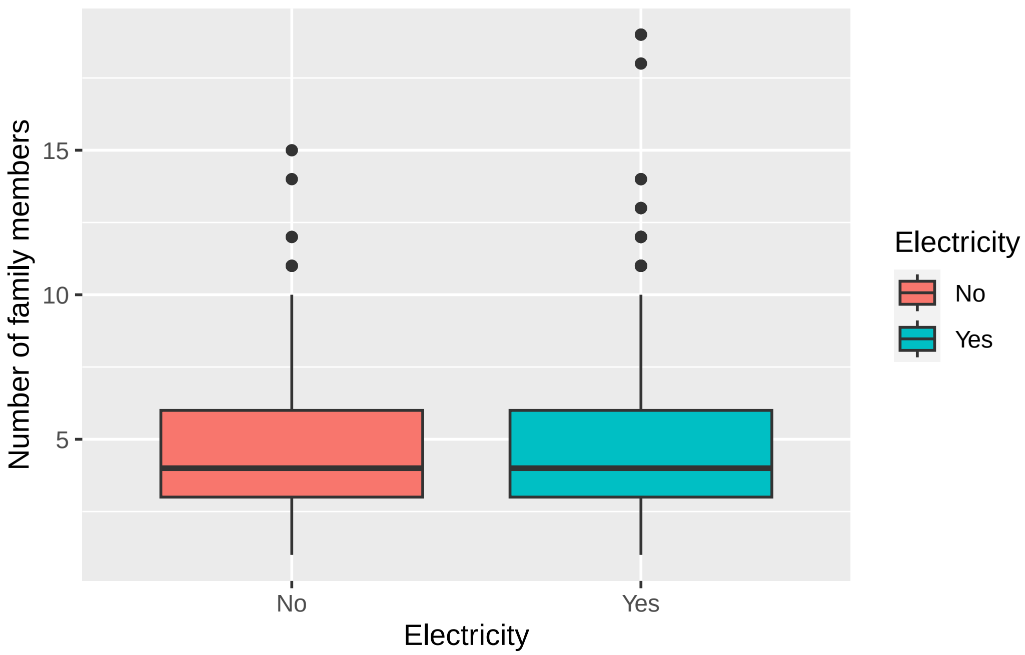


Figure 14: Number of family members by electricity

From the above Figure 14, households with electricity and without electricity have the same distribution of family members.

4 Formal analysis

4.1 Poisson Regression Model

4.1.1 Fit model with all variables

```
# As the response variable is the number of people living in a household, which is counts
data$Total.Number.of.Family.members=as.numeric(as.character(
  data$Total.Number.of.Family.members))
data$Number.of.bedrooms=as.numeric(as.character(data$Number.of.bedrooms))
model1=glm(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,d
model1%>%
  summary()
```

```
Call:
glm(formula = 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, family = poisson, data = data)

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                              1.597e+00  6.095e-02
Total.Household.Income                                  -2.385e-07  5.634e-08
Total.Food.Expenditure                                   2.930e-06  1.880e-07
Household.Head.SexMale                                   2.631e-01  3.053e-02
Household.Head.Age                                      -3.797e-03  8.105e-04
Type.of.HouseholdSingle Family                          -3.467e-01  2.291e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.058e-01  1.809e-01
House.Floor.Area                                        -4.940e-04  3.402e-04
House.Age                                               -3.715e-03  1.030e-03
Number.of.bedrooms                                       5.011e-02  1.234e-02
ElectricityYes                                          -9.028e-02  2.850e-02
                                                        z value Pr(>|z|)
(Intercept)                                              26.210  < 2e-16 ***
Total.Household.Income                                   -4.234 2.29e-05 ***
Total.Food.Expenditure                                   15.588  < 2e-16 ***
Household.Head.SexMale                                    8.616  < 2e-16 ***
Household.Head.Age                                       -4.684 2.81e-06 ***
Type.of.HouseholdSingle Family                          -15.135  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.585 0.558423
House.Floor.Area                                         -1.452 0.146476
House.Age                                                -3.606 0.000311 ***
Number.of.bedrooms                                        4.061 4.89e-05 ***
ElectricityYes                                           -3.168 0.001536 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1551.8  on 2111  degrees of freedom
AIC: 8511.9

Number of Fisher Scoring iterations: 5
```

```r
confint(model1)%>%
  kable()
```

	2.5 %	97.5 %
(Intercept)	1.4777012	1.7166106
Total.Household.Income	-0.0000004	-0.0000001
Total.Food.Expenditure	0.0000026	0.0000033
Household.Head.SexMale	0.2036003	0.3232971
Household.Head.Age	-0.0053862	-0.0022092
Type.of.HouseholdSingle Family	-0.3915529	-0.3017466
Type.of.HouseholdTwo or More Nonrelated Persons/Members	-0.4820181	0.2294578
House.Floor.Area	-0.0011694	0.0001642
House.Age	-0.0057424	-0.0017039
Number.of.bedrooms	0.0259109	0.0742825
ElectricityYes	-0.1458759	-0.0341516

```r
levels(data$Household.Head.Sex)
```

```
[1] "Female" "Male"
```

```r
levels(data$Type.of.Household)
```

```
[1] "Extended Family"
[2] "Single Family"
[3] "Two or More Nonrelated Persons/Members"
```

```r
levels(data$Electricity)
```

```
[1] "No"  "Yes"
```

The default baseline in R being taken as the one which comes first alphabetically. So these three categorical variables adopt female, Extended Family, No as baseline.

From the above summary we can observe that one continuous explanatory variable floor area is not significant and compared to extended family, Two or More Nonrelated Persons/Members is not significant while single family is significant according to the p-value and the 95% CI of estimates of coefficients.

4.1.1.1 Rate Ratio

```
model_summary <- summary(model1)
coef <- model_summary$coefficients[,1]
std_err <- model_summary$coefficients[,2]
rate_ratio <- exp(model_summary$coef)
conf_interval <- exp(cbind(coef - 1.96 * std_err, coef + 1.96 * std_err))
result <- data.frame(coef = coef, std_err = std_err, rate_ratio = rate_ratio, conf_interva
print(result)
```

```
                                                                coef
(Intercept)                                              1.597427e+00
Total.Household.Income                                  -2.385545e-07
Total.Food.Expenditure                                   2.930463e-06
Household.Head.SexMale                                   2.630600e-01
Household.Head.Age                                      -3.796566e-03
Type.of.HouseholdSingle Family                          -3.467288e-01
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.058474e-01
House.Floor.Area                                        -4.940074e-04
House.Age                                               -3.714620e-03
Number.of.bedrooms                                       5.011218e-02
ElectricityYes                                          -9.028251e-02
                                                            std_err
(Intercept)                                             6.094682e-02
Total.Household.Income                                  5.634096e-08
Total.Food.Expenditure                                  1.879964e-07
Household.Head.SexMale                                  3.053305e-02
Household.Head.Age                                      8.104823e-04
Type.of.HouseholdSingle Family                          2.290952e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members 1.808782e-01
House.Floor.Area                                        3.402039e-04
House.Age                                               1.030238e-03
Number.of.bedrooms                                      1.233996e-02
ElectricityYes                                          2.849982e-02
                                                        rate_ratio.Estimate
(Intercept)                                                       4.9403037
Total.Household.Income                                            0.9999998
Total.Food.Expenditure                                            1.0000029
Household.Head.SexMale                                            1.3009048
Household.Head.Age                                                0.9962106
Type.of.HouseholdSingle Family                                    0.7069970
Type.of.HouseholdTwo or More Nonrelated Persons/Members           0.8995620
```

```
House.Floor.Area                                                0.9995061
House.Age                                                       0.9962923
Number.of.bedrooms                                              1.0513890
ElectricityYes                                                  0.9136730
                                                     rate_ratio.Std..Error
(Intercept)                                                       1.062842
Total.Household.Income                                            1.000000
Total.Food.Expenditure                                            1.000000
Household.Head.SexMale                                            1.031004
Household.Head.Age                                                1.000811
Type.of.HouseholdSingle Family                                    1.023174
Type.of.HouseholdTwo or More Nonrelated Persons/Members           1.198269
House.Floor.Area                                                  1.000340
House.Age                                                         1.001031
Number.of.bedrooms                                                1.012416
ElectricityYes                                                    1.028910
                                                     rate_ratio.z.value
(Intercept)                                                2.415094e+11
Total.Household.Income                                     1.449251e-02
Total.Food.Expenditure                                     5.884700e+06
Household.Head.SexMale                                     5.516954e+03
Household.Head.Age                                         9.238931e-03
Type.of.HouseholdSingle Family                             2.673506e-07
Type.of.HouseholdTwo or More Nonrelated Persons/Members    5.570024e-01
House.Floor.Area                                           2.340801e-01
House.Age                                                  2.717131e-02
Number.of.bedrooms                                         5.803047e+01
ElectricityYes                                             4.209497e-02
                                                     rate_ratio.Pr...z..
(Intercept)                                                     1.000000
Total.Household.Income                                          1.000023
Total.Food.Expenditure                                          1.000000
Household.Head.SexMale                                          1.000000
Household.Head.Age                                              1.000003
Type.of.HouseholdSingle Family                                  1.000000
Type.of.HouseholdTwo or More Nonrelated Persons/Members         1.747914
House.Floor.Area                                                1.157747
House.Age                                                       1.000311
Number.of.bedrooms                                              1.000049
ElectricityYes                                                  1.001537
                                                            X1        X2
(Intercept)                                          4.3840416 5.5671462
Total.Household.Income                               0.9999997 0.9999999
Total.Food.Expenditure                                  1.0000026 1.0000033
Household.Head.SexMale                                  1.2253361 1.3811338
Household.Head.Age                                      0.9946294 0.9977944
Type.of.HouseholdSingle Family                          0.6759532 0.7394666
Type.of.HouseholdTwo or More Nonrelated Persons/Members 0.6310510 1.2823238
House.Floor.Area                                        0.9988399 1.0001728
House.Age                                               0.9942825 0.9983061
Number.of.bedrooms                                      1.0262649 1.0771283
ElectricityYes                                          0.8640349 0.9661629
```

The result from the rate ratio agree with that from p-values and confidence intervals. We can observe that the type "Two or More Nonrelated Persons/Members" is not significantly different compared to the baseline "Extended family". So we can firstly merge these two kinds of types of household to "Not Single", while another is "Single Family".

```
data1=read.csv("dataset04.csv") # for conviniency, introduce a new dataset, which we can m
data1[data1$Type.of.Household!="Single Family",]$Type.of.Household="Not Single"
```

4.1.2 Fit model on the merged dataset

```
model2=glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Ho
model2%>%
  summary()
```

```
Call:
glm(formula = 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, family = poisson, data = data1)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     1.595e+00  6.081e-02  26.230  < 2e-16 ***
Total.Household.Income         -2.385e-07  5.633e-08  -4.233 2.30e-05 ***
Total.Food.Expenditure          2.933e-06  1.879e-07  15.611  < 2e-16 ***
Household.Head.SexMale          2.629e-01  3.053e-02   8.611  < 2e-16 ***
Household.Head.Age             -3.782e-03  8.102e-04  -4.668 3.04e-06 ***
Type.of.HouseholdSingle Family -3.454e-01  2.280e-02 -15.147  < 2e-16 ***
House.Floor.Area               -4.909e-04  3.401e-04  -1.443 0.148971    
House.Age                     -3.696e-03  1.030e-03  -3.589 0.000331 ***
Number.of.bedrooms             5.016e-02  1.234e-02   4.065 4.81e-05 ***
Electricity                   -9.036e-02  2.850e-02  -3.171 0.001521 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1552.2  on 2112  degrees of freedom
AIC: 8510.2

Number of Fisher Scoring iterations: 5
```

we can find the variable floor area is still not significant, so we remove it then.

4.1.3 Remove floor area

```
model3=glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Ho
```

```
model3%>%
  summary()
```

```
Call:
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income + 
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age + 
    Type.of.Household + House.Age + Number.of.bedrooms + Electricity, 
    family = poisson, data = data1)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     1.594e+00  6.080e-02  26.220  < 2e-16 ***
Total.Household.Income         -2.531e-07  5.537e-08  -4.570 4.87e-06 ***
Total.Food.Expenditure          2.937e-06  1.880e-07  15.622  < 2e-16 ***
Household.Head.SexMale          2.633e-01  3.053e-02   8.625  < 2e-16 ***
Household.Head.Age             -3.837e-03  8.093e-04  -4.741 2.12e-06 ***
Type.of.HouseholdSingle Family -3.458e-01  2.280e-02 -15.164  < 2e-16 ***
House.Age                      -3.742e-03  1.029e-03  -3.635 0.000278 ***
Number.of.bedrooms              4.454e-02  1.172e-02   3.802 0.000144 ***
```

```
Electricity                     -9.140e-02  2.849e-02  -3.209 0.001334 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1554.3  on 2113  degrees of freedom
AIC: 8510.4

Number of Fisher Scoring iterations: 5
```

The explanatory variables in final model are all significant , with an AIC value of 8510.4.

4.2 Overdispersion

```
ggplot(model2, aes(x=log(fitted(model2)), y=log((data$Total.Number.of.Family.members-fitte
geom_point(col="#f46d43") +
geom_abline(slope=1, intercept=0, col="#a6d96a", linewidth=1) +
ylab(expression((y-hat(mu))^2)) + xlab(expression(hat(mu)))
```

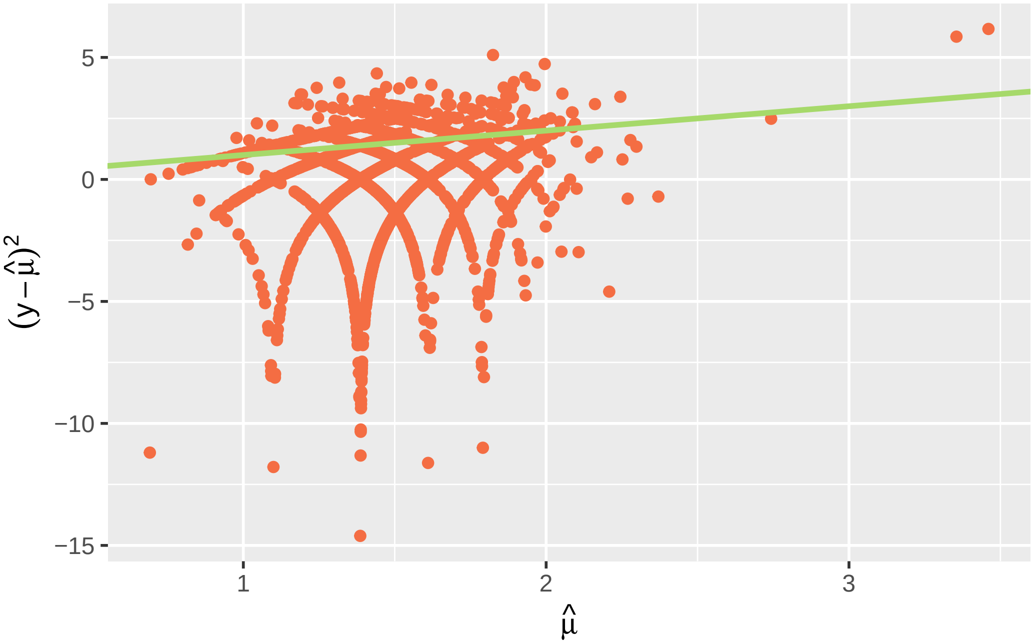


Figure 15: scatterplot of mean and variance

From Figure 15, we can find most of the points lie above the line of equality for mean and variance. In this case, we are not to able to determine which explanatory variables are significant.

4.2.1 Examine existence of dispersion

```
library(qcc)
qcc.overdispersion.test(data$Total.Number.of.Family.members)
```

```
Overdispersion test Obs.Var/Theor.Var Statistic   p-value
       poisson data          1.082586  2296.164 0.0042826
```

From the dispersion test we know that the p-value<0.05, indicating that the dispersion does exist in number of family members. So we should consider to fit a Quasi-Poisson model or a negative binomial model to the data.

4.2.2 Quasi-Poisson model

we can define a dispersion parameter ϕ such that $Var(Y_i) = \phi\mu_i$, we can estimate this parameter by

$$\hat{\phi} = \frac{X^2}{n-p}$$

```
X2=sum(resid(model1,type="pearson")^2)
dp=X2/model1$df.res
# With the use of the estimated dispersion parameter the Wald tests are not very reliable,
drop1(model1,test="F")
```

```
Single term deletions

Model:
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
                       Df Deviance    AIC  F value    Pr(>F)
<none>                      1551.8 8511.9
Total.Household.Income  1   1570.8 8528.8  25.7704 4.182e-07 ***
Total.Food.Expenditure  1   1737.1 8695.2 252.0856 < 2.2e-16 ***
Household.Head.Sex      1   1630.4 8588.4 106.8233 < 2.2e-16 ***
Household.Head.Age      1   1573.8 8531.9  29.9530 4.952e-08 ***
Type.of.Household       2   1774.8 8730.9 151.6907 < 2.2e-16 ***
House.Floor.Area        1   1554.0 8512.0   2.9244 0.0873964 .
House.Age               1   1565.0 8523.1  17.9624 2.350e-05 ***
Number.of.bedrooms      1   1568.3 8526.3  22.3752 2.391e-06 ***
Electricity             1   1561.7 8519.8  13.4388 0.0002526 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the model summary above, we are supposed to delete the variable House.Floor.Area.

```
model_quasi <- glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expend
summary(model_quasi)
```

```
Call:
glm(formula = 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, family = quasipoisson(link = "log"), data = data)

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                              1.597e+00  5.280e-02
Total.Household.Income                                  -2.385e-07  4.881e-08
Total.Food.Expenditure                                   2.930e-06  1.629e-07
Household.Head.SexMale                                   2.631e-01  2.645e-02
Household.Head.Age                                      -3.797e-03  7.021e-04
Type.of.HouseholdSingle Family                          -3.467e-01  1.985e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.058e-01  1.567e-01
House.Floor.Area                                        -4.940e-04  2.947e-04
House.Age                                               -3.715e-03  8.925e-04
Number.of.bedrooms                                       5.011e-02  1.069e-02
ElectricityYes                                          -9.028e-02  2.469e-02
                                                        t value Pr(>|t|)
(Intercept)                                              30.256  < 2e-16 ***
Total.Household.Income                                   -4.888 1.10e-06 ***
Total.Food.Expenditure                                   17.994  < 2e-16 ***
Household.Head.SexMale                                    9.946  < 2e-16 ***
Household.Head.Age                                       -5.407 7.11e-08 ***
Type.of.HouseholdSingle Family                          -17.471  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.676 0.499417
House.Floor.Area                                         -1.676 0.093836 .
House.Age                                                -4.162 3.28e-05 ***
Number.of.bedrooms                                        4.688 2.94e-06 ***
ElectricityYes                                           -3.657 0.000262 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.7504229)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1551.8  on 2111  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

In a Quasi-Poisson model, Two or More Nonrelated Persons/Members is still not significantly different compared to Extended Family. So we need to fit this model again using merged dataset.

```
model_quasi_1 <- glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expe
summary(model_quasi_1)
```

```
Call:
glm(formula = 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, family = quasipoisson(link = "log"), data = data1)

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.595e+00  5.267e-02  30.281  < 2e-16 ***
Total.Household.Income         -2.385e-07  4.880e-08  -4.887 1.10e-06 ***
Total.Food.Expenditure          2.933e-06  1.627e-07  18.022  < 2e-16 ***
Household.Head.SexMale          2.629e-01  2.645e-02   9.941  < 2e-16 ***
Household.Head.Age             -3.782e-03  7.018e-04  -5.389 7.89e-08 ***
Type.of.HouseholdSingle Family -3.454e-01  1.975e-02 -17.486  < 2e-16 ***
House.Floor.Area               -4.909e-04  2.946e-04  -1.666 0.095849 .  
House.Age                      -3.696e-03  8.920e-04  -4.144 3.55e-05 ***
Number.of.bedrooms              5.016e-02  1.069e-02   4.693 2.87e-06 ***
Electricity                    -9.036e-02  2.469e-02  -3.660 0.000258 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.7503333)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1552.2  on 2112  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Then we need to remove the floor area variable.

```
model_quasi_2 <- glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expe
summary(model_quasi_2)
```

```
Call:
```

```
glm(formula = Total.Number.of.Family.members ~ Total.Household.Income +
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age +
    Type.of.Household + House.Age + Number.of.bedrooms + Electricity,
    family = quasipoisson(link = "log"), data = data1)

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.594e+00  5.269e-02  30.252  < 2e-16 ***
Total.Household.Income         -2.531e-07  4.799e-08  -5.273 1.48e-07 ***
Total.Food.Expenditure          2.937e-06  1.630e-07  18.024  < 2e-16 ***
Household.Head.SexMale          2.633e-01  2.646e-02   9.951  < 2e-16 ***
Household.Head.Age             -3.837e-03  7.014e-04  -5.470 5.02e-08 ***
Type.of.HouseholdSingle Family -3.458e-01  1.976e-02 -17.496  < 2e-16 ***
House.Age                      -3.742e-03  8.922e-04  -4.194 2.85e-05 ***
Number.of.bedrooms              4.454e-02  1.015e-02   4.386 1.21e-05 ***
Electricity                    -9.140e-02  2.469e-02  -3.702 0.000219 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.7511975)

    Null deviance: 2217.8  on 2121  degrees of freedom
Residual deviance: 1554.3  on 2113  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Using the Quasi-Poisson model, we reach the same conclusion as what we get in the ordinary glm model, which removes only floor area variable.

4.2.3 Negative binomial models

Considering the Overdispersion, another choice is the Negative-binomial model.

```
model_nb=glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditu
summary(model_nb)
```

```
Call:
glm.nb(formula = 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, data = data, init.theta = 76069.16879, link = log)

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                             1.597e+00  6.095e-02
Total.Household.Income                                 -2.386e-07  5.634e-08
Total.Food.Expenditure                                  2.931e-06  1.880e-07
Household.Head.SexMale                                  2.631e-01  3.053e-02
Household.Head.Age                                     -3.797e-03  8.105e-04
Type.of.HouseholdSingle Family                         -3.467e-01  2.291e-02
Type.of.HouseholdTwo or More Nonrelated Persons/Members -1.058e-01  1.809e-01
House.Floor.Area                                       -4.940e-04  3.402e-04
House.Age                                              -3.715e-03  1.030e-03
Number.of.bedrooms                                      5.011e-02  1.234e-02
ElectricityYes                                         -9.029e-02  2.850e-02
                                                        z value Pr(>|z|)
(Intercept)                                              26.209  < 2e-16 ***
Total.Household.Income                                   -4.234 2.29e-05 ***
Total.Food.Expenditure                                   15.588  < 2e-16 ***
Household.Head.SexMale                                    8.615  < 2e-16 ***
Household.Head.Age                                       -4.684 2.81e-06 ***
Type.of.HouseholdSingle Family                          -15.134  < 2e-16 ***
Type.of.HouseholdTwo or More Nonrelated Persons/Members  -0.585 0.558455
House.Floor.Area                                         -1.452 0.146465
House.Age                                                -3.605 0.000312 ***
Number.of.bedrooms                                        4.061 4.89e-05 ***
ElectricityYes                                           -3.168 0.001536 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(76069.32) family taken to be 1)

    Null deviance: 2217.7  on 2121  degrees of freedom
Residual deviance: 1551.7  on 2111  degrees of freedom
AIC: 8513.9

Number of Fisher Scoring iterations: 1

              Theta:  76069
          Std. Err.:  280723
Warning while fitting theta: alternation limit reached
```

```
 2 x log-likelihood:  -8489.906
```

Similarly, we can see that the categorical variable Type.of.Household(Two or More Nonrelated Persons/Members) and continuous variable House.Floor.Area seem not to be statistically significant with the response variable.

```
model_nb1 <- glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expen
summary(model_nb1)
```

```
Call:
glm.nb(formula = 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, data = data1, init.theta = 75964.50263, link = log)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     1.595e+00  6.081e-02  26.229  < 2e-16 ***
Total.Household.Income         -2.385e-07  5.633e-08  -4.233 2.30e-05 ***
Total.Food.Expenditure          2.933e-06  1.879e-07  15.611  < 2e-16 ***
Household.Head.SexMale          2.629e-01  3.053e-02   8.611  < 2e-16 ***
Household.Head.Age             -3.782e-03  8.102e-04  -4.668 3.05e-06 ***
Type.of.HouseholdSingle Family -3.454e-01  2.280e-02 -15.146  < 2e-16 ***
House.Floor.Area               -4.909e-04  3.401e-04  -1.443 0.148959    
House.Age                      -3.696e-03  1.030e-03  -3.589 0.000332 ***
Number.of.bedrooms              5.016e-02  1.234e-02   4.065 4.81e-05 ***
Electricity                    -9.037e-02  2.850e-02  -3.171 0.001520 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(75964.42) family taken to be 1)

    Null deviance: 2217.7  on 2121  degrees of freedom
Residual deviance: 1552.1  on 2112  degrees of freedom
AIC: 8512.3

Number of Fisher Scoring iterations: 1


              Theta:  75965 
          Std. Err.:  280536 
Warning while fitting theta: alternation limit reached 
```

```
 2 x log-likelihood:  -8490.261
```

```
model_nb1$aic
```

```
[1] 8512.261
```

We firstly fit a negative model using the merged dataset and find the floor area is still not significant. So we need to remove it in our next model.

```
model_nb2 <- glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expen
summary(model_nb2)
```

```
Call:
glm.nb(formula = Total.Number.of.Family.members ~ Total.Household.Income + 
    Total.Food.Expenditure + Household.Head.Sex + Household.Head.Age + 
    Type.of.Household + House.Age + Number.of.bedrooms + Electricity, 
    data = data1, init.theta = 76018.70336, link = log)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     1.594e+00  6.080e-02  26.219  < 2e-16 ***
Total.Household.Income         -2.531e-07  5.537e-08  -4.570 4.87e-06 ***
Total.Food.Expenditure          2.937e-06  1.880e-07  15.622  < 2e-16 ***
Household.Head.SexMale          2.633e-01  3.053e-02   8.625  < 2e-16 ***
Household.Head.Age             -3.837e-03  8.093e-04  -4.741 2.12e-06 ***
Type.of.HouseholdSingle Family -3.458e-01  2.280e-02 -15.163  < 2e-16 ***
House.Age                      -3.742e-03  1.029e-03  -3.635 0.000278 ***
Number.of.bedrooms              4.454e-02  1.172e-02   3.801 0.000144 ***
Electricity                    -9.141e-02  2.849e-02  -3.209 0.001334 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(76018.77) family taken to be 1)

    Null deviance: 2217.7  on 2121  degrees of freedom
Residual deviance: 1554.2  on 2113  degrees of freedom
AIC: 8512.4
```

```
Number of Fisher Scoring iterations: 1

              Theta:  76019
          Std. Err.:  280041
Warning while fitting theta: alternation limit reached

 2 x log-likelihood:  -8492.384
```

```
model_nb2$aic
```

```
[1] 8512.384
```

Using the negative binomial model, all the variables except floor area are significant and the AIC value is 8512.384.

4.3 Using AIC to choose best model

4.3.1 GLM model

```
c(glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househo
glm(Total.Number.of.Family.members~Total.Food.Expenditure+Household.Head.Sex+Household.Hea
glm(Total.Number.of.Family.members~Total.Household.Income+Household.Head.Sex+Household.Hea
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
glm(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Household
```

```
[1] 8510.362 8530.664 8694.374 8587.067 8530.923 8731.536 8521.788 8522.761
[9] 8518.495
```

Removed variable	AIC value
None	8510.362
Total.Household.Income	8530.664
Total.Food.Expenditure	8694.374
Household.Head.Sex	8587.067

Removed variable	AIC value
Household.Head.Age	8530.923
Type.of.Household	8731.536
House.Age	8521.788
Number.of.bedrooms	8522.761
Electricity	8518.495

In ordinary GLM model, the full model with all explanatory variables except House.Floor.Area has the lowest AIC value.

4.3.2 Negative binomial model

```
c(glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Hous
glm.nb(Total.Number.of.Family.members~Total.Food.Expenditure+Household.Head.Sex+Household.
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Household.Head.Sex+Household.
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
glm.nb(Total.Number.of.Family.members~Total.Household.Income+Total.Food.Expenditure+Househ
```

```
[1] 8512.384 8532.686 8696.396 8589.086 8532.944 8733.542 8523.809 8524.781
[9] 8520.517
```

Removed variable	AIC value
None	8512.384
Total.Household.Income	8532.686
Total.Food.Expenditure	8696.396
Household.Head.Sex	8589.086
Household.Head.Age	8532.944
Type.of.Household	8733.542
House.Age	8523.809
Number.of.bedrooms	8524.781
Electricity	8520.517

5 Final model

We find that GLM model with only floor area variable removed has the lowest AIC value.

The final model is:

$$log(Total.Number.of.Family.members) = \beta_0 + \beta_1 \cdot Total.Household.Income + \beta_2 \cdot Total.Food.Expenditure + \beta_3 \cdot$$

$$\mathbb{1}_{\text{Male}}(x) = \begin{cases} 1 & \text{if the head of household is Male,} \\ 0 & \text{if the head of household is female.} \end{cases}$$

$$\mathbb{1}_{\text{Family}}(x) = \begin{cases} 1 & \text{Single family,} \\ 0 & \text{Not Single Family.} \end{cases}$$

$$\mathbb{1}_{\text{Electricity}}(x) = \begin{cases} 1 & \text{if the house has electricity,} \\ 0 & \text{Otherwise.} \end{cases}$$

For extended family and two or more nonrelated persons/members, the final model is:

$$log(Total.Number.of.Family.members) = 1.594 - 2.531 \times 10^{-7} \cdot Total.Household.Income + 2.937 \times 10^{-6} \cdot Total.F$$

For single family, the final model is:

$$log(Total.Number.of.Family.members) = 1.25 - 2.531 \times 10^{-7} \cdot Total.Household.Income + 2.937 \times 10^{-6} \cdot Total.Fo$$

6 Conclusion

After removing the insignificant variables and comparing the AIC values of different models, it is found that the variables Total.Household.Income, Total.Food.Expenditure, Household.Head.Sex, Household.Head.Age, Type.of.Household, House.Age, Number.of.bedrooms and Electricity could influence response variable Total.Number.of.Family.members (the number of people living in a household).