Group_2_Analysis

Group 2

2022/3/20

```
library(tidyverse)
library(moderndive)
library(gapminder)
library(sjPlot)
library(stats)
library(jtools)
library(MASS)
library(kableExtra)
library(olsrr)
library(qcc)
library(skimr)
library(GGally)
```

```
#import data
data<-read.csv("dataset2.csv")

#processing discrete data
data[, 4] <- as.factor(data[, 4])
data[, 6] <- as.factor(data[, 6])
data[, 11] <- as.factor(data[, 11])

data = data[, -2]
```

Introduction

Exploratory Data Analysis

Statistics Summary

In order to make it more clear, we devide all variables into continous and discrete variables.

```
#divide continuous and discrete variables
con_var = c("Total.Number.of.Family.members","Total.Household.Income",
            "Total.Food.Expenditure","Household.Head.Age",
           "House.Floor.Area", "House.Age", "Number.of.bedrooms")

dis_var = c("Household.Head.Sex", "Type.of.Household", "Electricity")
```

Table 1: Summary statistics of continuous variables

Variable	Mean	SD	Min	Median	Max	IQR
Total.Number.of.Family.members	4	2	1	4	16	2
Total.Household.Income	216685	263207	18784	140483	2891788	89919
Total.Food.Expenditure	70760	41638	10488	62590	413844	24118
Household.Head.Age	51	14	15	51	87	10
House.Floor.Area	49	49	5	36	750	24
House.Age	16	13	0	14	105	8
Number.of.bedrooms	2	1	0	2	7	0

Table 2: Summary statistics of discrete variables

Variable	Counts
Household.Head.Sex	Mal: 983, Fem: 266
Type.of.Household	Sin: 900, Ext: 344, Two: 5
Electricity	1: 1069, 0: 180

```r
# data summary of continuous variables
data[con_var] %>%
  skim() %>%
  transmute(Variable=skim_variable,
            Mean=round(numeric.mean),
            SD=round(numeric.sd),
            Min=numeric.p0,
            Median=numeric.p50,
            Max=numeric.p100,
            IQR = numeric.p75-numeric.p50) %>%
  kable(caption = '\\label{tab:summaries_con} Summary statistics of continuous variables') #%>%
```

```r
  #kable_styling(font_size = 10, latex_options = "hold_position")
```

From Table 1, we can see that the median of "Total.Household.Income" is much less than its mean. It indicates that high income families have considerable impact on the samples. However, when we look at "Total.Food.Expenditure", the difference of mean and median is much smaller comparing with "Total.Household.Income".

The range of "Total.Number.of.Family.members" is from 1 to 16. The median and the mean are equal which is 4. The standard deviation is 2 so the variance is 4 which is equal to the mean. This means the distribution of "Total.Number.of.Family.members" may be poisson distribution.

Regarding other continuous variables, we do not find something special.

```r
# data summary of discrete variables
data[dis_var] %>%
  skim() %>%
   transmute(Variable=skim_variable,
            Counts=factor.top_counts) %>%
  kable(caption = '\\label{tab:summaries_dis} Summary statistics of discrete variables') #%>%
  #kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 2 shows that most house heads are male, there are a few of families consist of unrelated persons and some houses still do not have electricity.

Data Visualization Analysis

```
#plot the distribution of "Total.Number.of.Family.members"
ggplot(data, aes(x=Total.Number.of.Family.members)) +
  geom_histogram()
```

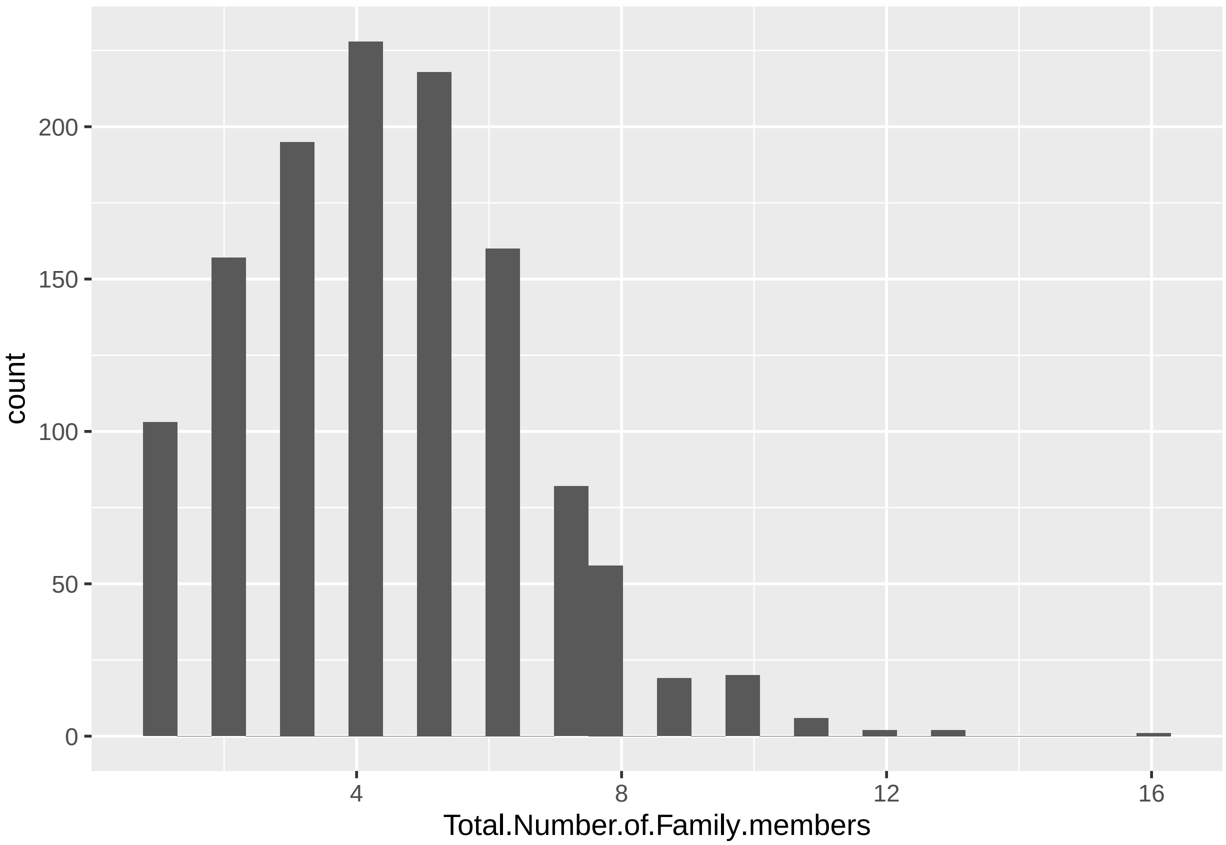

Figure 1: Distribution of Total.Number.of.Family.members

As Figure 1 shows that the variance of "Total.Number.of.Family.members" is equal to its mean. In addition, Figure 1 shows that the distribution of "Total.Number.of.Family.members" is symmetry from 1 to 9 which is similar to poisson distribution. In general, we can infer that except some families with people numbers higher than 9, the distribution of "Total.Number.of.Family.members" is likely to be poisson distribution and poisson logistic regression may be applied to solve our problem.

```
#boxplot of "household.Head.Sex" vs "Total.Number.of.Family.members"
ggplot(data, aes(x=Household.Head.Sex, y=Total.Number.of.Family.members)) +
  geom_boxplot()
```

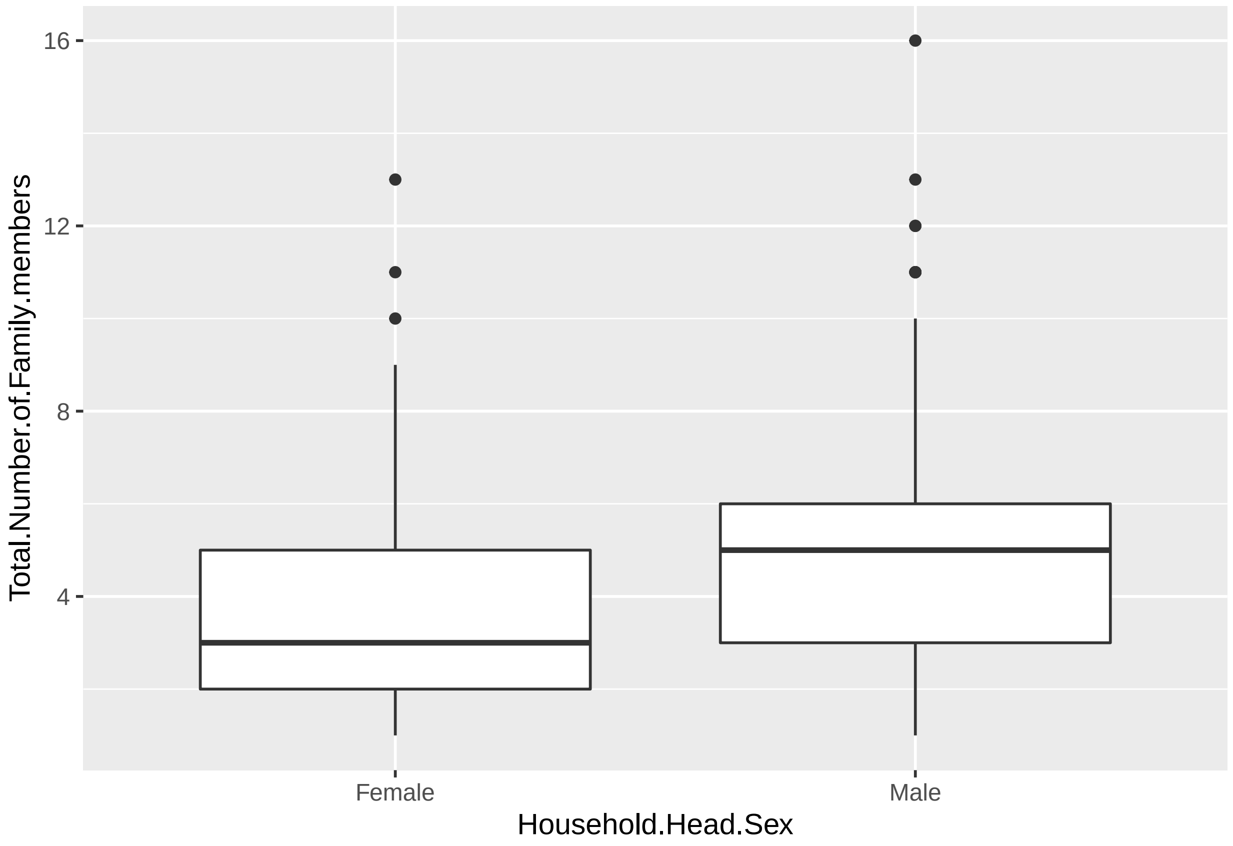

Figure 2: Boxplot of Household.Head.Sex and Total.Number.of.Family.members

From Figure 2, we can see that on average, male head families have more members than female. However, we also can see there is overlap in the IQR's.

```
#boxplot of "Type.of.Household" vs "Total.Number.of.Family.members"
ggplot(data, aes(x=Type.of.Household, y=Total.Number.of.Family.members)) +
  geom_boxplot()
```

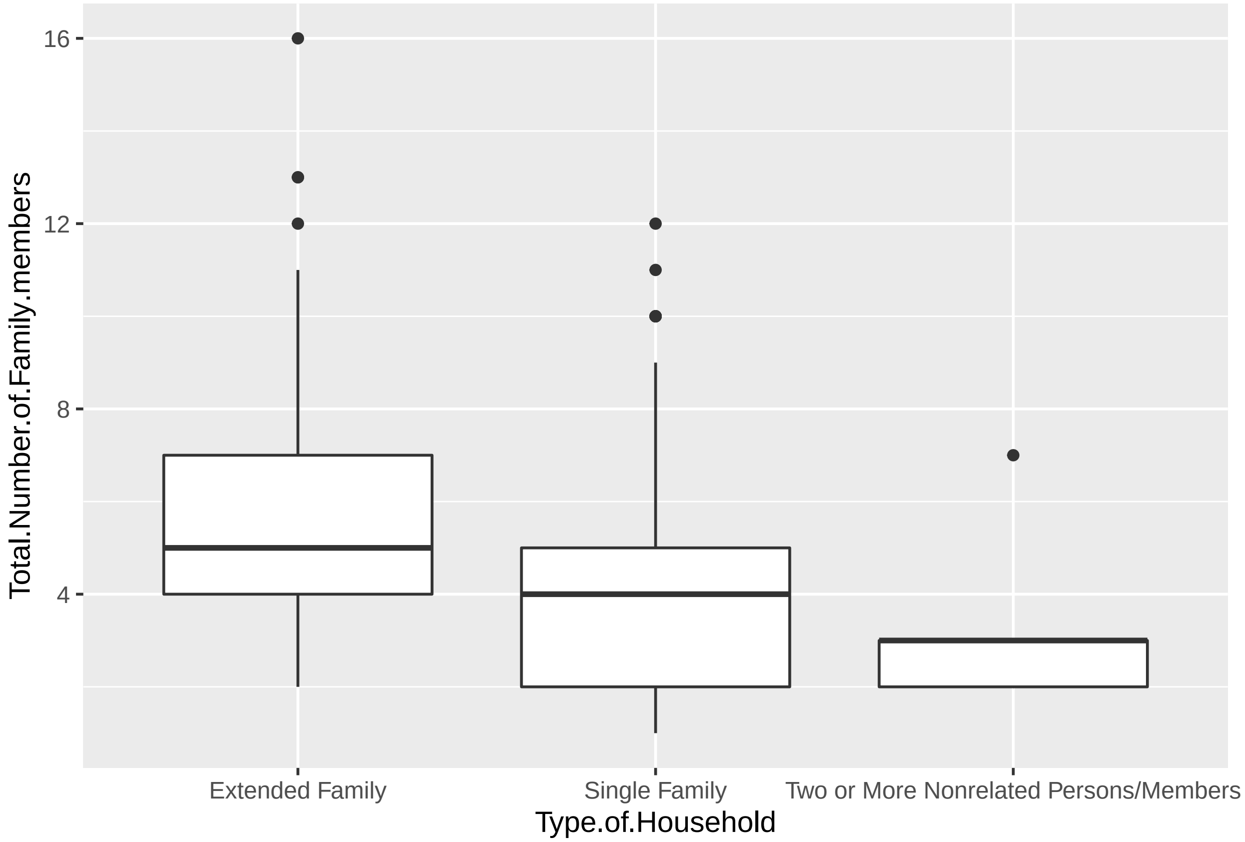

Figure 3: Boxplot of Type.of.Household and Total.Number.of.Family.members

From Figure 3, we can see that on average, different types of household, on average, have different numbers of members as well. Extended Families have more members than single families and single families have more members than nonrelated persons families. There are overlaps between them as well.

```
#boxplot of "Electricity" vs "Total.Number.of.Family.members"
ggplot(data, aes(x=as.factor(Electricity), y=Total.Number.of.Family.members)) +
  geom_boxplot()
```

Figure 4 shows that on average, electricity has no impact on the numbers of members in each family.

```
#plot matrix of continuous variables
ggpairs(data[, con_var], aes(alpha = 0.1))
```

From Figure 5, we can see that the correlation between "Total.Number.of.Family.members" and other continuous variables weak and very weak. When we look at the first column of this matrix, we can find that

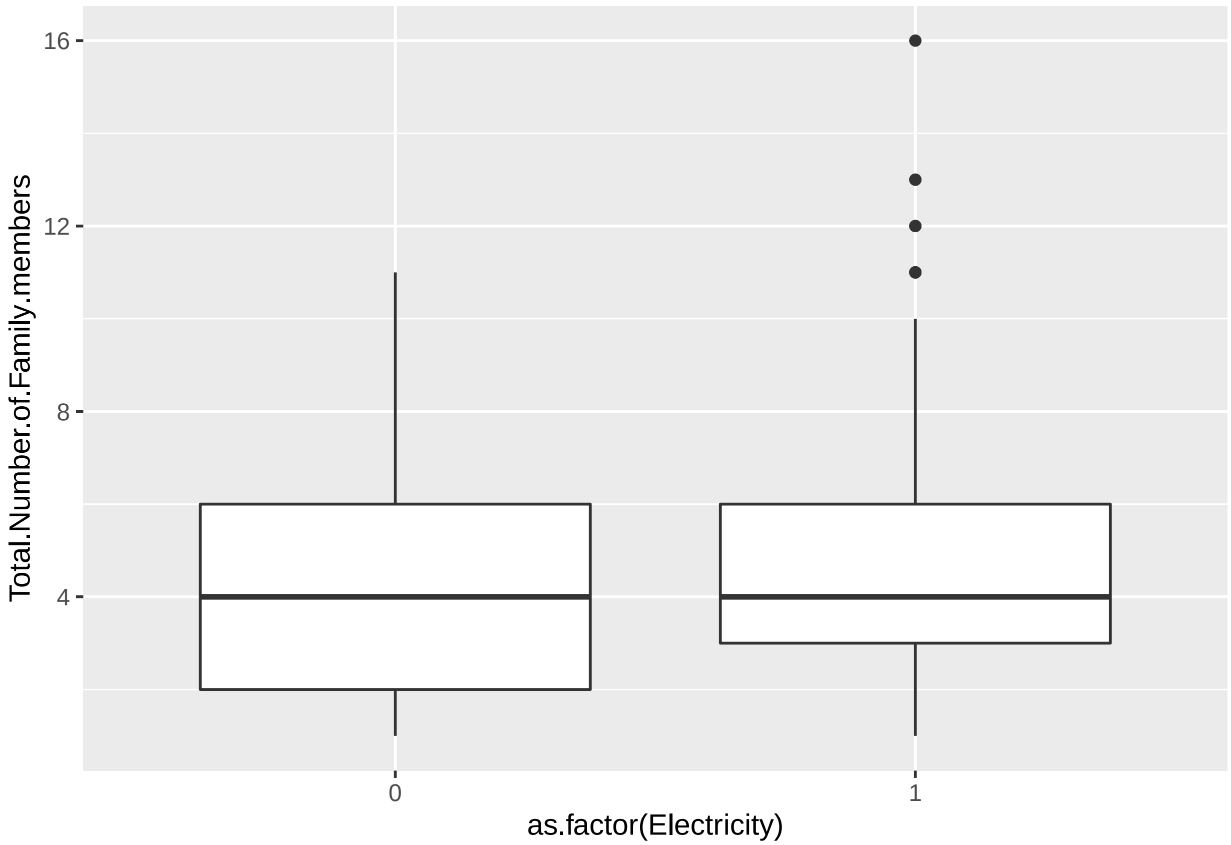

Figure 4: Boxplot of Electricity and Total.Number.of.Family.members

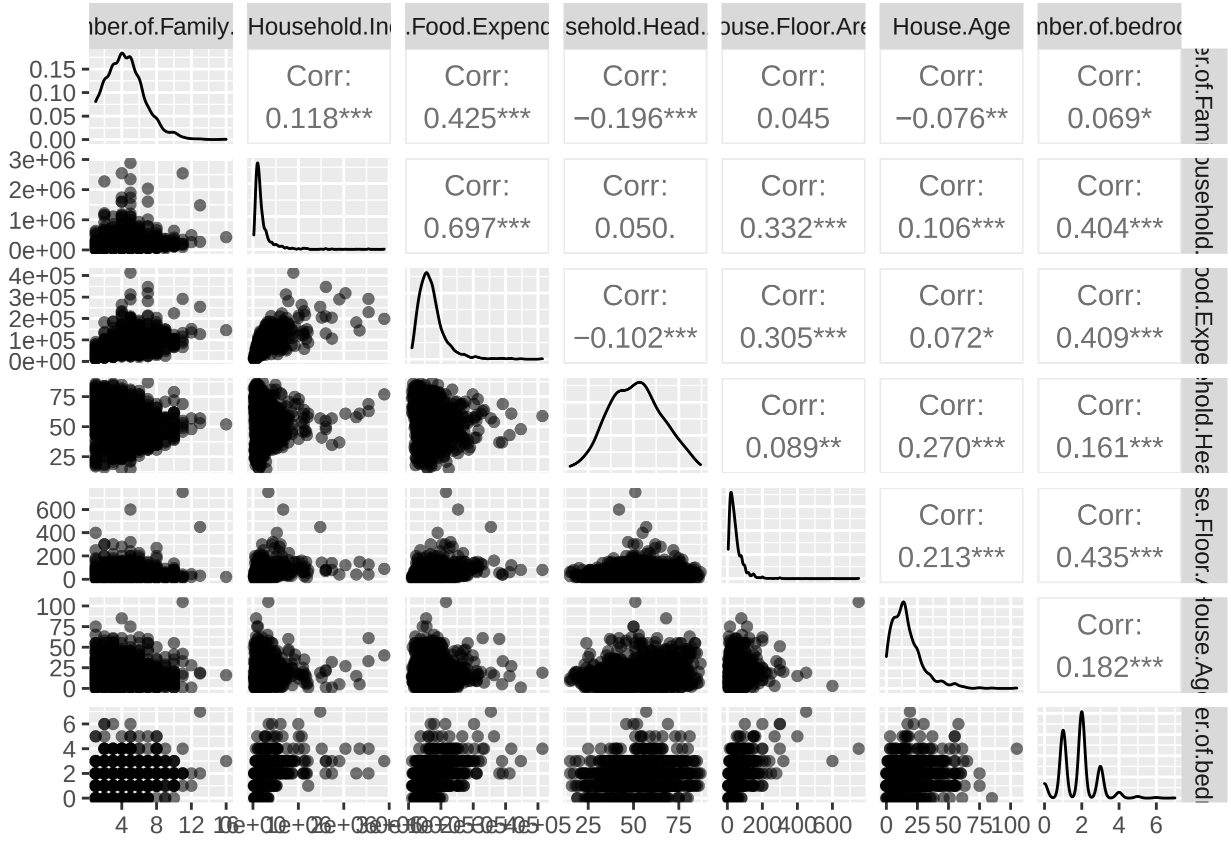

Figure 5: Matrix of plots of continuous variables

although “Total.Number.of.Family.members” and other continuous variables only have weak linear relationship, there may be non-linear relationship between them.

We we look at the correlation of other variables except “Total.Number.of.Family.members”, we can find that only “Total.Household.Income” and “Total.Food.Expenditure” have moderate positive correlation and others only have weak and very weak correlation.

Modelling and Results

Conclusions and Future Work