Group_01_Analysis.Rmd

Group_01

3/14/2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(moderndive)
library(skimr)
library(readr)
library(Stat2Data)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(broom)
library(sjPlot)
```

```
## Learn more about sjPlot with 'browseVignettes("sjPlot")'.
```

```
library(DescTools)
library(knitr)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr)
```

Introduction

Dataset comes 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. In this study, we are going to identify the most influential variables on the number of people living in a household using a Generalised Linear Model. Below you can see an overview of the data and variables

```
## Rows: 1,725
## Columns: 11
## $ Income           <int> 480332, 198235, 82785, 107589, 189322, 152883, 198621~
## $ Region           <fct> CAR, CAR, CAR, CAR, CAR, CAR, CAR, CAR, CAR, CAR, CAR~
## $ Expenditure      <int> 117848, 67766, 61609, 78189, 94625, 73326, 104644, 95~
## $ Gender           <fct> Female, Male, Male, Male, Male, Male, Male, Male, Fem~
## $ Age              <int> 49, 40, 39, 52, 65, 46, 45, 33, 17, 53, 49, 35, 38, 5~
## $ Type             <fct> Extended Family, Single Family, Single Family, Single~
## $ Number_of_Family <int> 4, 3, 6, 3, 4, 4, 5, 5, 2, 6, 4, 7, 7, 3, 2, 4, 5, 8,~
## $ Area             <int> 80, 42, 35, 30, 54, 40, 35, 35, 35, 70, 40, 35, 35, 5~
## $ HouseAge         <int> 75, 15, 12, 15, 16, 7, 18, 48, 8, 12, 9, 17, 5, 43, 7~
## $ Bedrooms         <int> 3, 2, 1, 1, 3, 2, 1, 2, 1, 3, 2, 3, 1, 3, 1, 1, 1, 1,~
## $ Electricity      <int> 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,~
```

Income is the annual household income (in Philippine peso) *Region* is the region of the Philippines which the data came from *Expenditure* is the annual expenditure by the household on food (in Philippine peso)
Gender is the head of the households sex
Age is the head of the households age (in years)
Type is the relationship between the group of people living in the house
$Number_of Family$ is the number of people living in the house
Area is the floor area of the house (in m^2)
$HouseAge$ is the age of the building (in years)
bedrooms is the number of bedrooms in the house
Electricity indicates that if the house have electricity? (1=Yes, 0=No)

```
#Region column removed from the data
data <- data %>% select(-2)

# The third category of the type variable (Two or More Nonrelated Persons/Members) which only 8 observa

data <- data %>% filter(Type != "Two or More Nonrelated Persons/Members") %>%
  mutate(Type = factor(Type, levels = c("Single Family", "Extended Family")))
```

```
#You can run the analysis with 3 categories of the type included by commenting the lines above an inste

# data <- data %>% mutate(Type = as.character(Type),
#   Type=replace(Type, Type=="Two or More Nonrelated Persons/Members", "Others")) %>%
#   mutate(Type = factor(Type, levels = c("Single Family", "Extended Family","Others")))


glimpse(data)
```

```
## Rows: 1,717
## Columns: 10
## $ Income           <int> 480332, 198235, 82785, 107589, 189322, 152883, 198621~
## $ Expenditure      <int> 117848, 67766, 61609, 78189, 94625, 73326, 104644, 95~
## $ Gender           <fct> Female, Male, Male, Male, Male, Male, Male, Male, Fem~
## $ Age              <int> 49, 40, 39, 52, 65, 46, 45, 33, 17, 53, 49, 35, 38, 5~
## $ Type             <fct> Extended Family, Single Family, Single Family, Single~
## $ Number_of_Family <int> 4, 3, 6, 3, 4, 4, 5, 5, 2, 6, 4, 7, 7, 3, 2, 4, 5, 8,~
## $ Area             <int> 80, 42, 35, 30, 54, 40, 35, 35, 35, 70, 40, 35, 35, 5~
## $ HouseAge         <int> 75, 15, 12, 15, 16, 7, 18, 48, 8, 12, 9, 17, 5, 43, 7~
## $ Bedrooms         <int> 3, 2, 1, 1, 3, 2, 1, 2, 1, 3, 2, 3, 1, 3, 1, 1, 1, 1,~
## $ Electricity      <int> 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,~
```

Exploratory Data Analysis

The following tables and graphs are produced to provide statistical summaries and graphs to see the distribution variables and their relationship and identify any possible outliers.

The variable "Region" is removed as it has only one level and will not contribute to the upcoming analysis as it has only one state. Furthermore, the **Type** variable was initially composed of three levels of "Extended Family", "Single Family" and "Two or More Nonrelated Persons/Members" categories. There were only 8 observations(less than 0.05 percent of total observations) in the last category. This category removed since it wasn't of significant importance.

Table 1: Summary Statistics

Variable	Min	Mean	SD	Median	Max
Income	11988	269524.8	275079.4	188050	6042860
Expenditure	6781	80249.4	41241.7	73459	327724
Age	17	52.2	14.5	52	99
Number_of_Family	1	4.7	2.3	4	15
Area	5	90.9	99.3	54	900
HouseAge	0	22.9	15.3	20	100
Bedrooms	0	2.3	1.4	2	9
Electricity	0	0.9	0.3	1	1

Referring to 1, the difference between mean and median indicates the presence of outliers in the data. The following pair plots fuhrer support the presence of outliers.

Figure 3 presents the distribution of outcome variables with regards to each of the continuous variables. The boxplots can help us identify the outliers

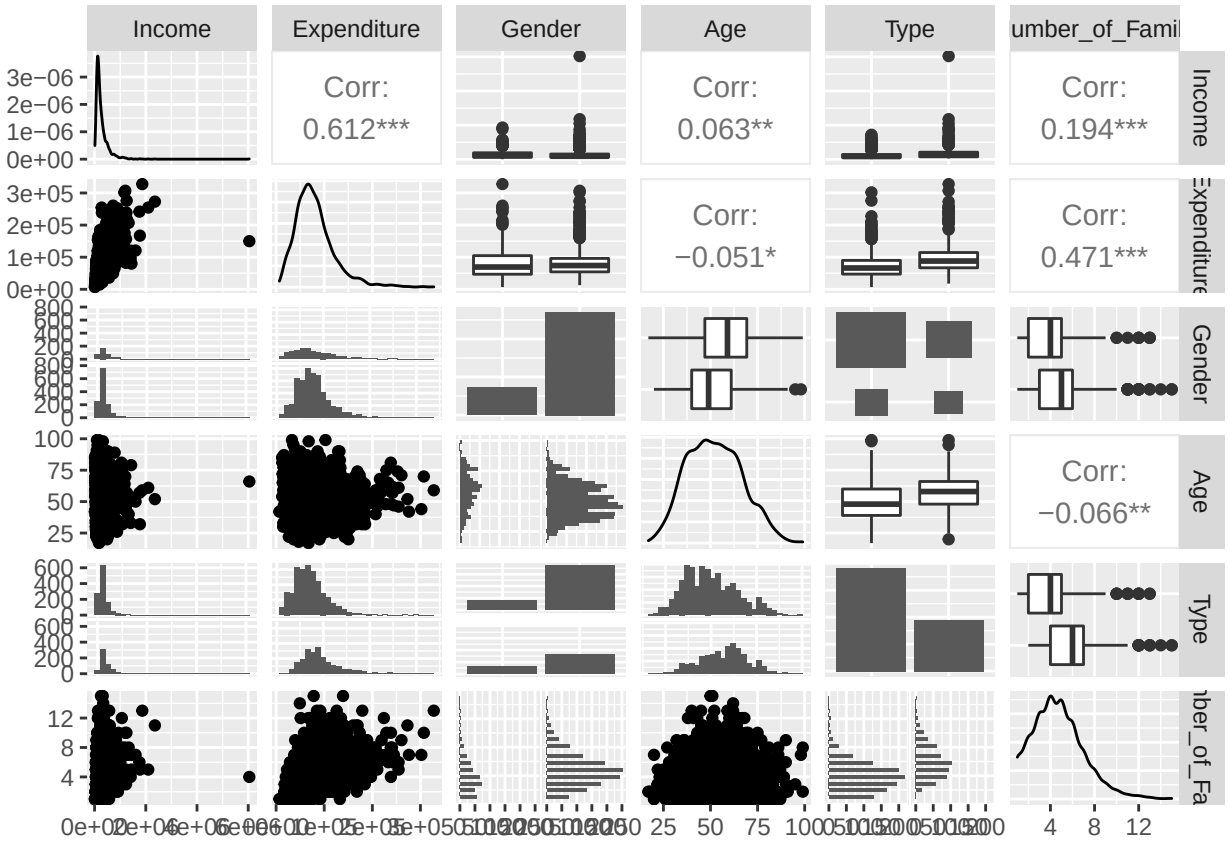


Figure 1: The pair plot (first five variables) shows the relationship between each of the two variables

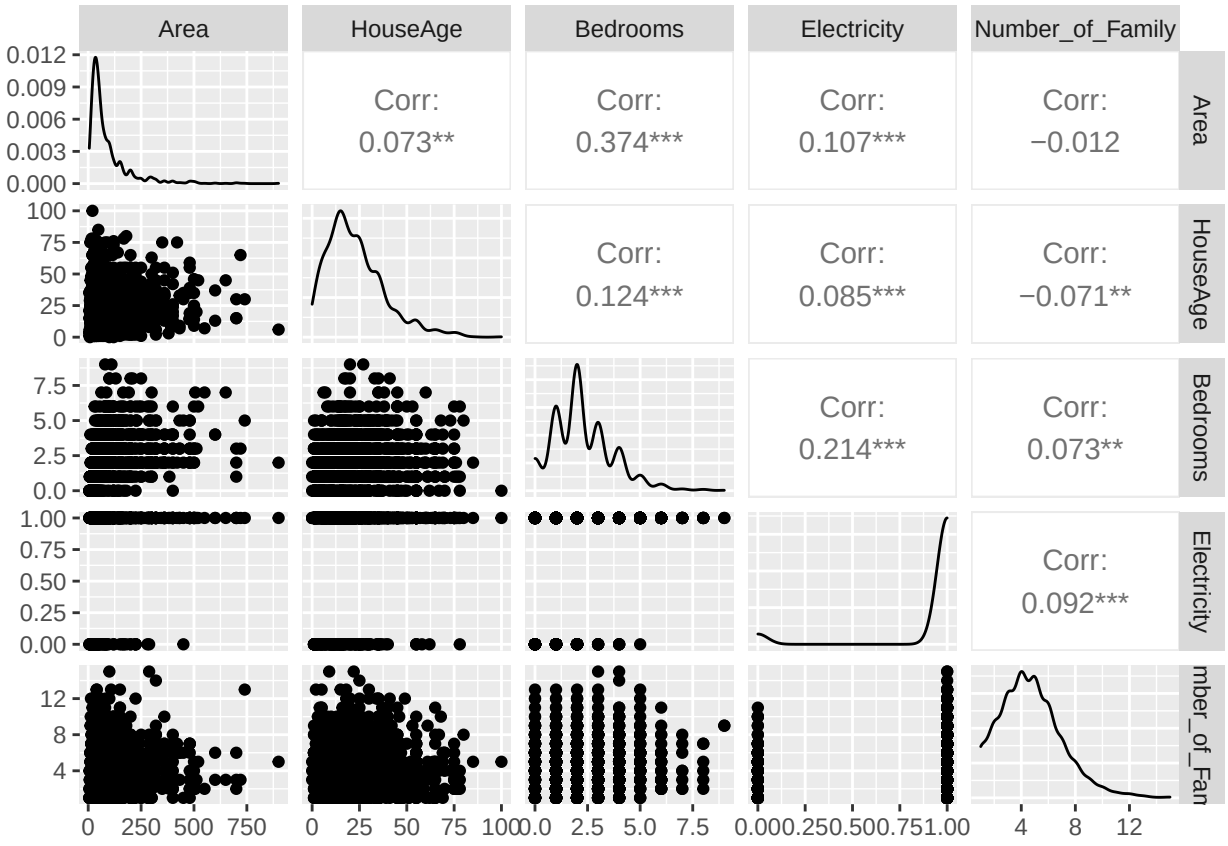


Figure 2: The pair plot (second five variables) shows the relationship between each of the two variables

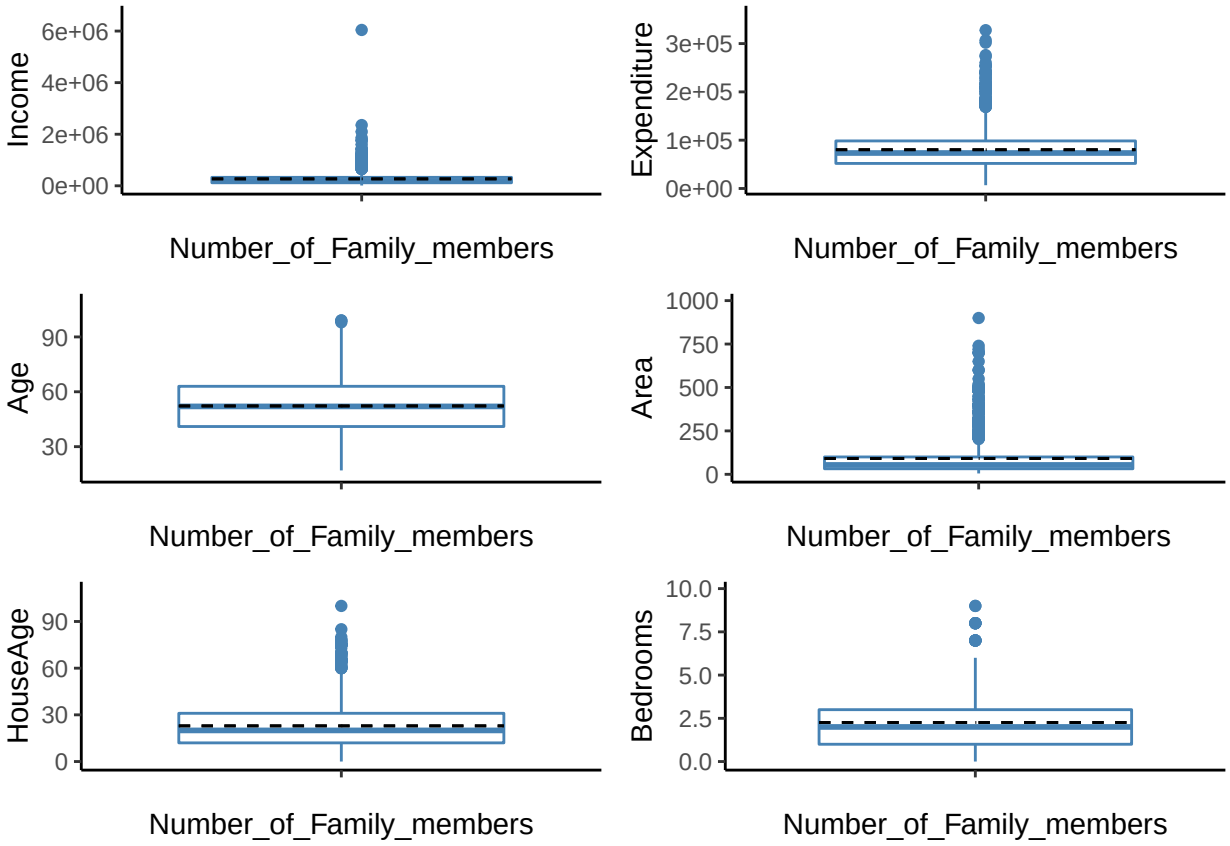


Figure 3: The boxplot of the outcome variables vs each of the continuous explanatory variables

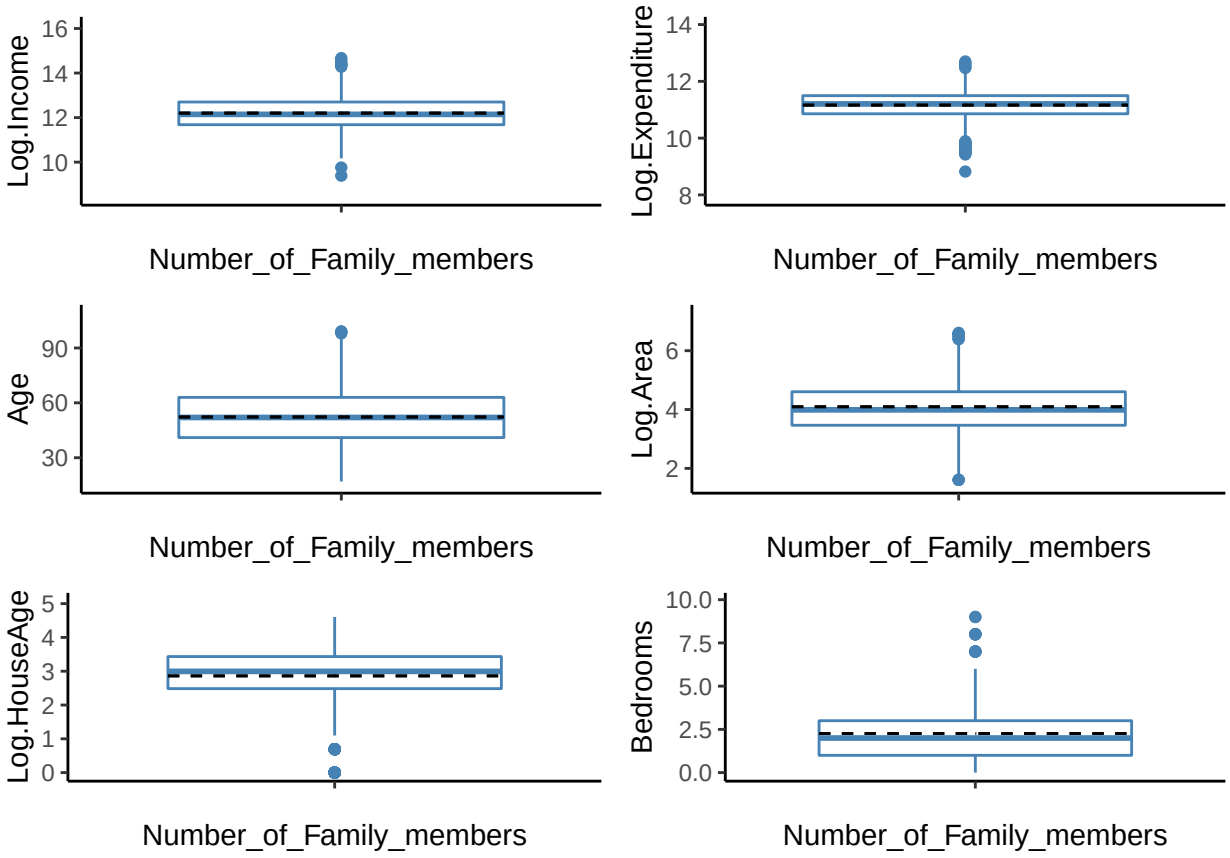


Figure 4: The boxplot of the outcome variables vs each of the continuous explanatory variables after removing outliers and applying log transformation

Once the outliers are spotted and removed, the skewness in data is decreased. The Figure 4 displays the box plots after removing outliers and log transforming Income, Expenditure and Area.

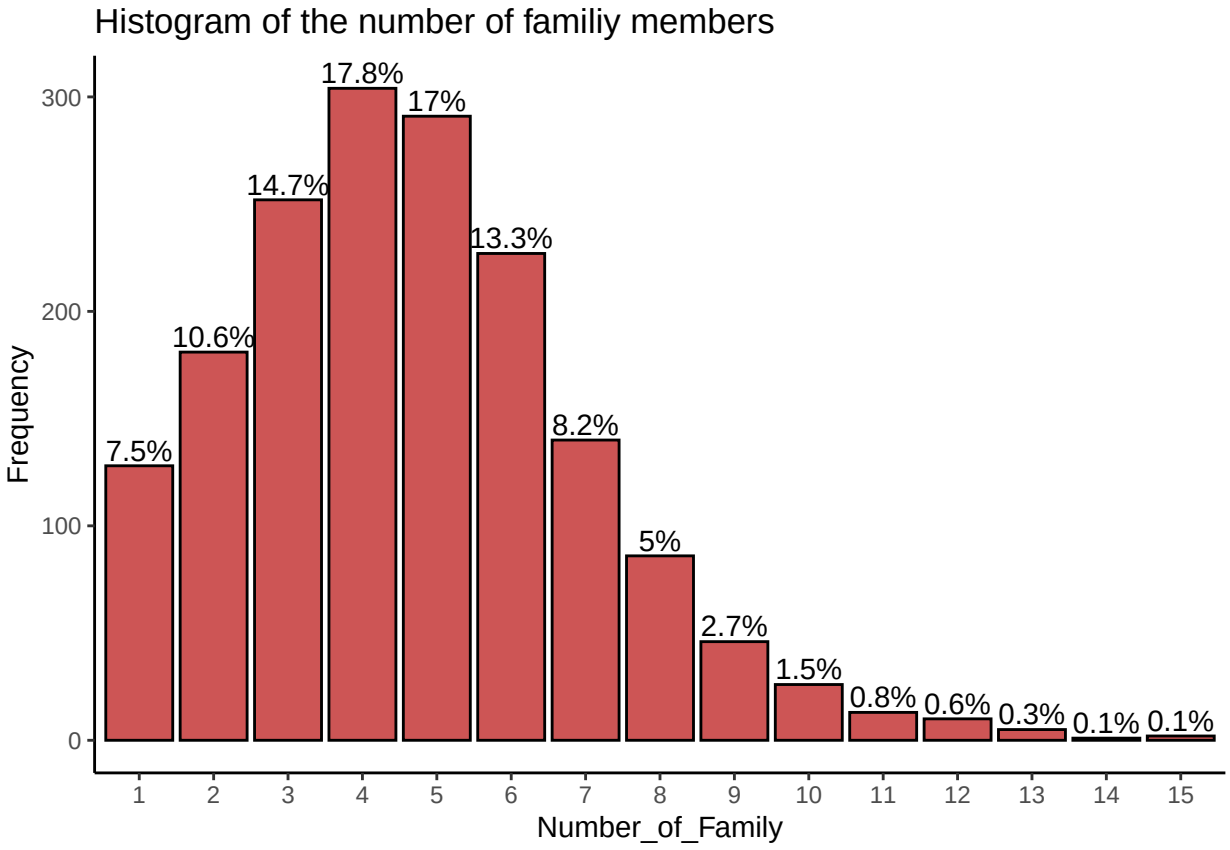
```
## TableGrob (3 x 2) "arrange": 6 grobs
##   z     cells    name           grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
## 3 3 (2-2,1-1) arrange gtable[layout]
## 4 4 (2-2,2-2) arrange gtable[layout]
## 5 5 (3-3,1-1) arrange gtable[layout]
## 6 6 (3-3,2-2) arrange gtable[layout]
```

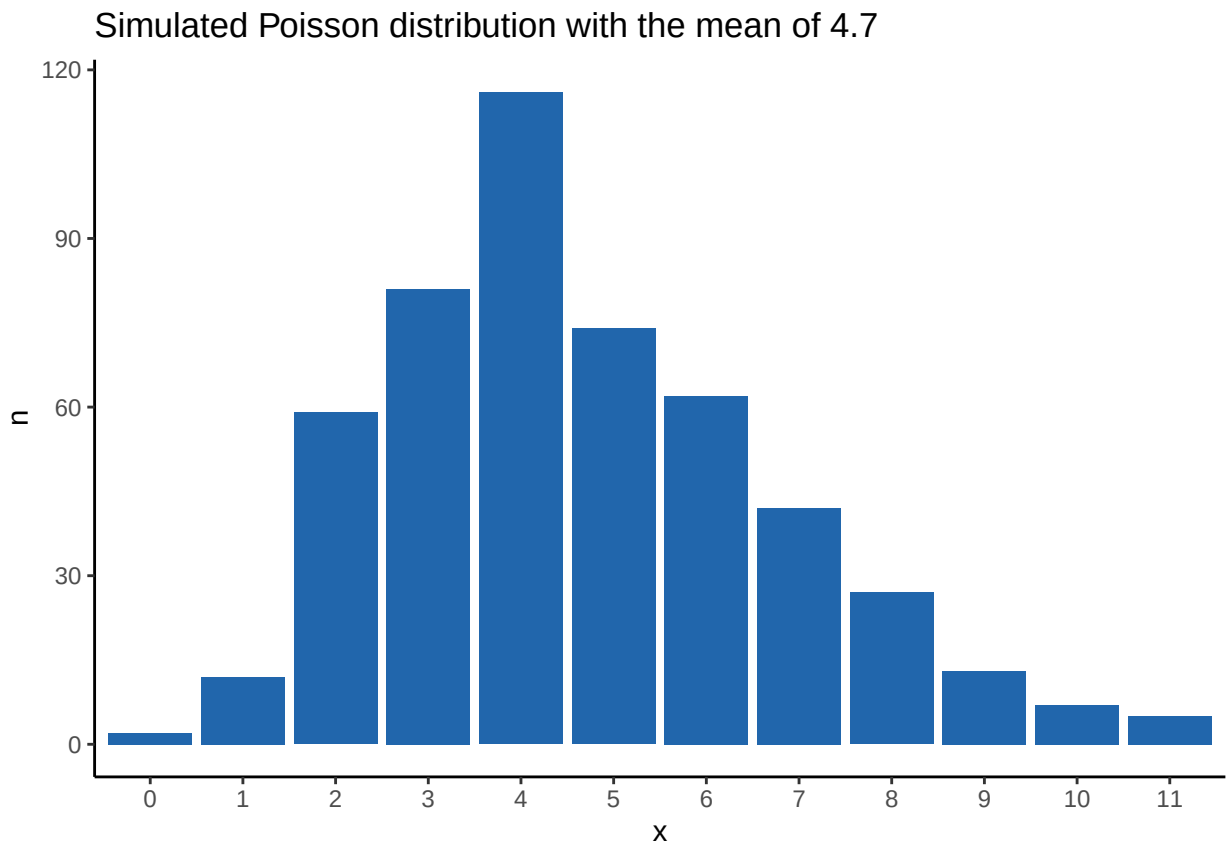
The table 2 shows the summary statistics after removing outliers and transforming the variables of Income, Expenditure and Aare using log transformation. The difference between medians and means are now narrower.

Table 2: Summary Statistics after outliers removed

Variable	Min	Mean	SD	Median	Max
Log.Income	9.4	12.2	0.7	12.1	14.7
Log.Expenditure	8.8	11.2	0.5	11.2	12.7
Age	17.0	52.3	14.5	52.0	99.0
Number_of_Family	1.0	4.7	2.3	4.0	15.0
Log.Area	1.6	4.1	0.9	4.0	6.6
Log.HouseAge	0.0	2.9	0.8	3.0	4.6
Bedrooms	0.0	2.3	1.4	2.0	9.0
Electricity	0.0	0.9	0.3	1.0	1.0

The Figure **??** and **??** shows the distribution of household sizes and simulated Poisson distribution with the mean value of 4.7 respectively.


Histogram of the number of familiy members
Frequency
Number_of_Family
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
7.5%
10.6%
14.7%
17.8%
17%
13.3%
8.2%
5%
2.7%
1.5%
0.8%
0.6%
0.3%
0.1%
0.1%



The histogram (Figure **??**) of the family size resembles a Poisson distribution.

Formal Data Analysis

In this section we model the data to identify the most influential factors on the number of family members using Poisson distribution since the the outcome variabl is a count data. Then the goodness of fit comparison is made to select the best model based on AIC and Deviance. However, for the a Poisson to completly hold the variance and mean shall be equal. The variance and the mean are 5.4 and 4.7. We can say that they are roughly equal.

Poisson model

```
model.poisson = glm(data = data, Number_of_Family ~ Log.Income + Log.Expenditure +
                      Gender + Age + Type + Log.Area + Log.HouseAge + Bedrooms + Electricity,
                    family = poisson(link = "log"))
initial.poisson.AIC = model.poisson$aic
summary(model.poisson)
```

```
##
## Call:
## glm(formula = Number_of_Family ~ Log.Income + Log.Expenditure +
##     Gender + Age + Type + Log.Area + Log.HouseAge + Bedrooms +
```

```
##     Electricity, family = poisson(link = "log"), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4411  -0.6375  -0.1069   0.4520   3.9304
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -3.8696965  0.3029732 -12.772  < 2e-16 ***
## Log.Income          -0.1062227  0.0255279  -4.161 3.17e-05 ***
## Log.Expenditure      0.5983754  0.0352401  16.980  < 2e-16 ***
## GenderMale           0.1876184  0.0298720   6.281 3.37e-10 ***
## Age                 -0.0013076  0.0008877  -1.473   0.1407
## TypeExtended Family  0.3018733  0.0250289  12.061  < 2e-16 ***
## Log.Area            -0.0224880  0.0157892  -1.424   0.1544
## Log.HouseAge        -0.0181523  0.0137208  -1.323   0.1858
## Bedrooms            -0.0176744  0.0101908  -1.734   0.0829 .
## Electricity         -0.0252561  0.0480667  -0.525   0.5993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2009.7  on 1711  degrees of freedom
## Residual deviance: 1174.3  on 1702  degrees of freedom
## AIC: 6813.8
##
## Number of Fisher Scoring iterations: 4
```

We now try to see if we can improve model AIC and decrease the deviance using step function. In this method we start from the full model and every time drop one variable and calculate the AIC. This procedure is continued until no further reduction in AIC is observed.

```
## Start:  AIC=6813.83
## Number_of_Family ~ Log.Income + Log.Expenditure + Gender + Age +
##     Type + Log.Area + Log.HouseAge + Bedrooms + Electricity
##
##                   Df Deviance    AIC
## - Electricity      1   1174.6 6812.1
## - Log.HouseAge     1   1176.1 6813.6
## <none>                 1174.3 6813.8
## - Log.Area         1   1176.4 6813.9
## - Age              1   1176.5 6814.0
## - Bedrooms         1   1177.3 6814.8
## - Log.Income       1   1191.8 6829.3
## - Gender           1   1215.2 6852.7
## - Type             1   1318.2 6955.7
## - Log.Expenditure  1   1470.3 7107.8
##
## Step:  AIC=6812.1
## Number_of_Family ~ Log.Income + Log.Expenditure + Gender + Age +
##     Type + Log.Area + Log.HouseAge + Bedrooms
##
##                   Df Deviance    AIC
```

```
## - Log.HouseAge     1   1176.5 6812.0
## <none>                 1174.6 6812.1
## - Log.Area         1   1176.7 6812.2
## - Age              1   1176.7 6812.2
## - Bedrooms         1   1177.7 6813.2
## - Log.Income       1   1192.4 6827.9
## - Gender           1   1215.5 6851.0
## - Type             1   1318.2 6953.7
## - Log.Expenditure  1   1470.8 7106.3
##
## Step:  AIC=6812
## Number_of_Family ~ Log.Income + Log.Expenditure + Gender + Age +
##     Type + Log.Area + Bedrooms
##
##                   Df Deviance    AIC
## <none>                 1176.5 6812.0
## - Log.Area         1   1178.7 6812.2
## - Age              1   1179.2 6812.7
## - Bedrooms         1   1180.3 6813.8
## - Log.Income       1   1194.2 6827.7
## - Gender           1   1218.7 6852.2
## - Type             1   1319.1 6952.6
## - Log.Expenditure  1   1473.6 7107.1

##
## Call:  glm(formula = Number_of_Family ~ Log.Income + Log.Expenditure +
##     Gender + Age + Type + Log.Area + Bedrooms, family = poisson(link = "log"),
##     data = data)
##
## Coefficients:
##         (Intercept)           Log.Income      Log.Expenditure
##           -3.919539            -0.106676              0.597847
##          GenderMale                  Age  TypeExtended Family
##            0.190261            -0.001442              0.300007
##            Log.Area             Bedrooms
##           -0.023387            -0.019717
##
## Degrees of Freedom: 1711 Total (i.e. Null);  1704 Residual
## Null Deviance:       2010
## Residual Deviance: 1176  AIC: 6812

##
## Call:
## glm(formula = Number_of_Family ~ Log.Income + Log.Expenditure +
##     Gender + Type + Bedrooms + Log.HouseAge, family = poisson(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4677  -0.6473  -0.1068   0.4611   3.8787
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.041709   0.287723 -14.047  < 2e-16 ***
```

```
## Log.Income          -0.117601   0.024866  -4.729 2.25e-06 ***
## Log.Expenditure      0.612225   0.034373  17.811  < 2e-16 ***
## GenderMale           0.194011   0.029562   6.563 5.28e-11 ***
## TypeExtended Family  0.289119   0.023970  12.062  < 2e-16 ***
## Bedrooms            -0.025173   0.009384  -2.682  0.00731 **
## Log.HouseAge        -0.021985   0.013516  -1.627  0.10383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2009.7  on 1711  degrees of freedom
## Residual deviance: 1178.8  on 1705  degrees of freedom
## AIC: 6812.3
##
## Number of Fisher Scoring iterations: 4
```

By removing Electricity and Area and Age, the AIC is reduced from 6813.8 to 6812.3. We now check for the goodness of fit by comparing it against the null model. The 95 percent χ^2(p = 0.95, df = 6) equals 12.6. Taking the difference in deviances (likelihood ratio test) results in the value of 830.8 which is significant when compared to 12.6. Therefore, there is no deviance of lack of fit with our model after removing the variables. The Log.HouseAge is not also significantly different from zero and by removing it, the AIC remains almost constant, however, the BIC is further reduced by 5 after removing Log.HouseAge. The Pseudo_R2 remains almost the same for all models. Table 3 presents AIC, BIC and Pseudo-R2 for the models assessed in our analysis. They are shown in order of dropping variables from the full model. For example "F—Age" represents the model with Electricity, Log.Area and Area removed. Given above, The formula below is proposed to model expected count:

$$log(Number\_of\_Family) = \beta_0 + \beta_1 \cdot log(Income) + \beta_2 \cdot log(FoodExpenditure) + \beta_3 \cdot Gender + \beta_4 \cdot Bedrooms + \beta_5 \cdot Type$$

Table 3: Model comparison values for different models.

Model	AIC	BIC	pseudo_R2
Fullmodel	6813.8	6868.3	0.109
F-Electricity	6812.1	6861.1	0.109
F–Log.Area	6812.2	6855.8	0.109
F—Age	6812.3	6850.5	0.109
F—-Log.House	6813.0	6845.6	0.109

Parameter estimates

Table 4 displays the parameter estimates. The main explanatory variables are significantly different from zero

Table 4: parameter estimates of the Poisson regression model

Parameter	Estimate	Std. Error	Lower CI	Upper CI	P.value
(Intercept)	-4.125	0.283	-4.681	-3.572	0.000
Log.Income	-0.118	0.025	-0.167	-0.069	0.000
Log.Expenditure	0.614	0.034	0.547	0.682	0.000
GenderMale	0.198	0.029	0.141	0.256	0.000

Parameter	Estimate	Std. Error	Lower CI	Upper CI	P.value
TypeExtended Family	0.286	0.024	0.239	0.333	0.000
Bedrooms	-0.028	0.009	-0.046	-0.010	0.003

The rate ratio are obtained by exponentiating the coefficients(Table 5). Figure **??** exhibits the rate ration for the different explantory variables

Table 5: Rate ratios 95 percent confidence interval

	Estimate	Lower CI	Upper CI
(Intercept)	0.02	0.01	0.03
Log.Income	0.89	0.85	0.93
Log.Expenditure	1.85	1.73	1.98
GenderMale	1.22	1.15	1.29
TypeExtended Family	1.33	1.27	1.39
Bedrooms	0.97	0.96	0.99

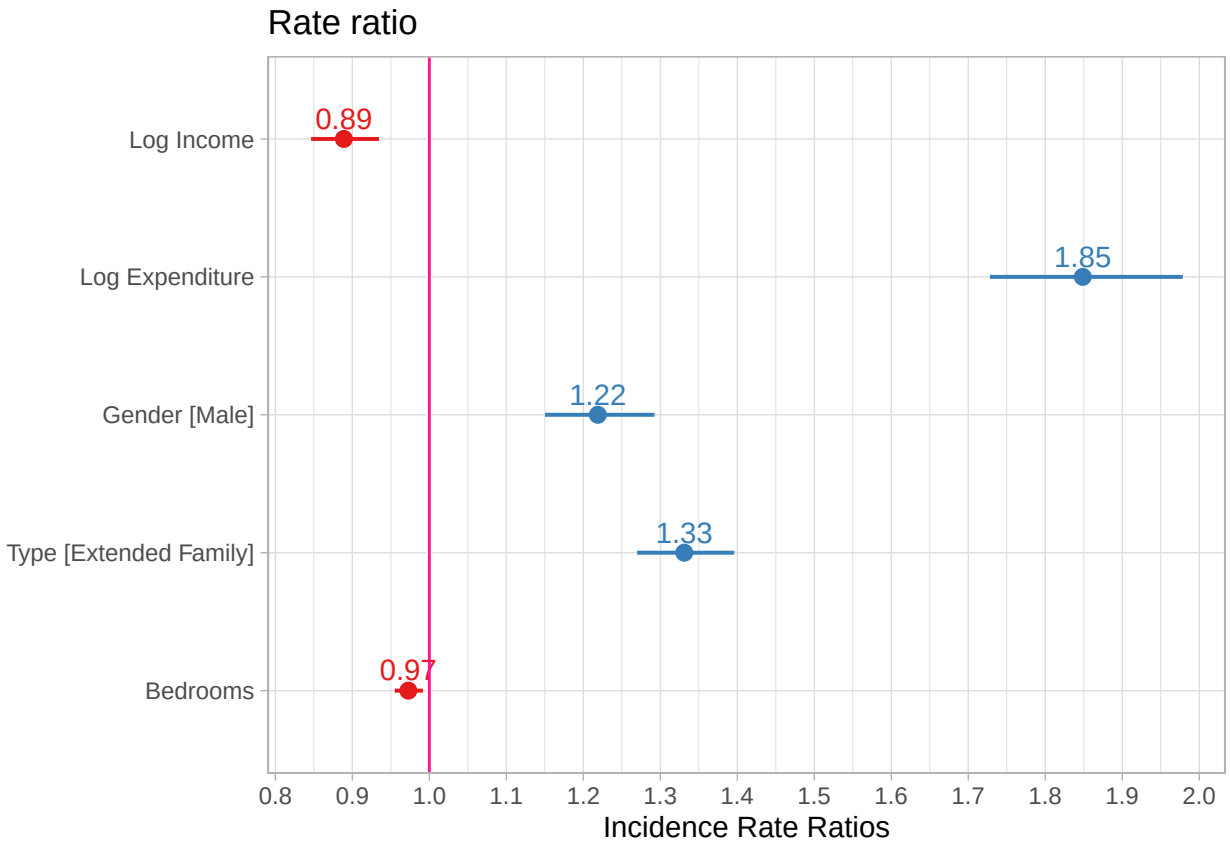


Figure 5 shows the relationship between each of the variables and the outcome variable

Figure 6 illustrates the predictions of the model. The difference between real values and predictions can be seen in Figure 7

```
## Warning: Removed 2 rows containing missing values (geom_col).
```

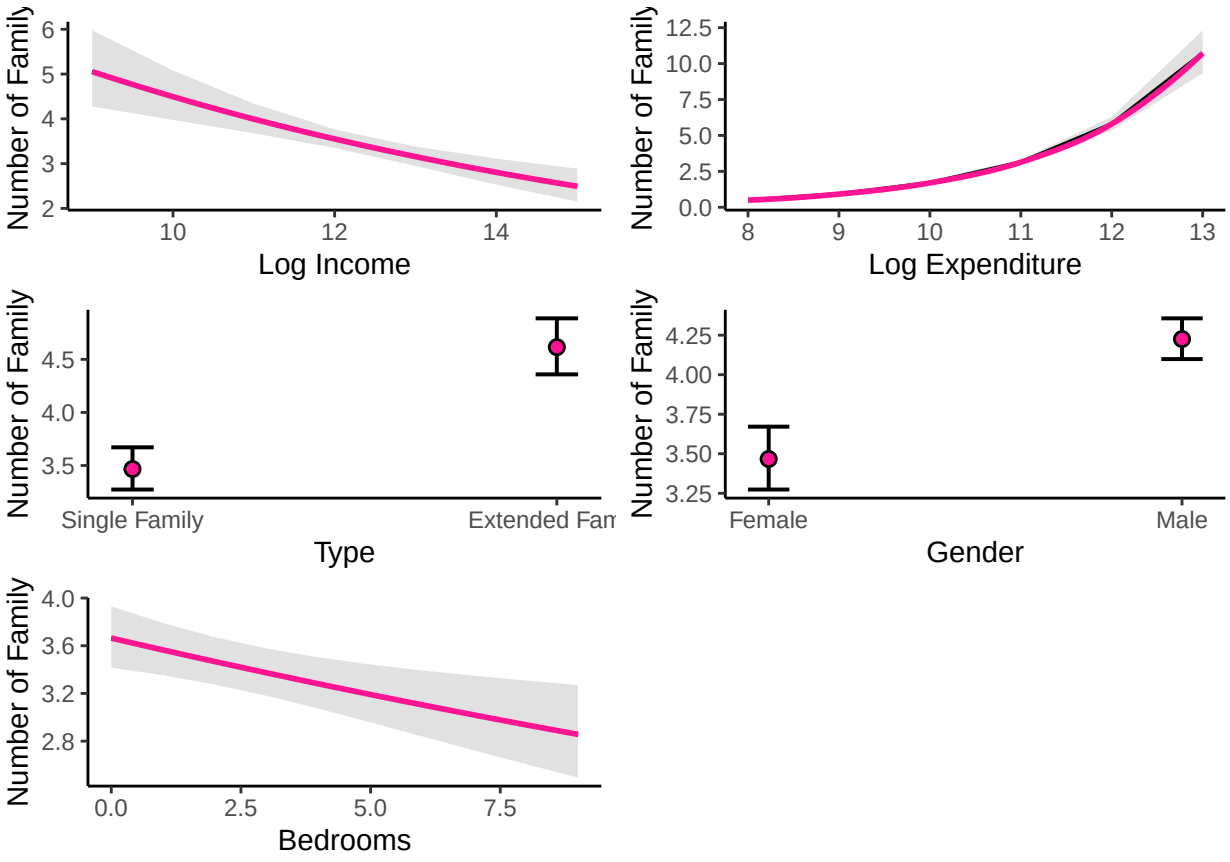


Figure 5: The relationship between explanatory variables and the number of family members

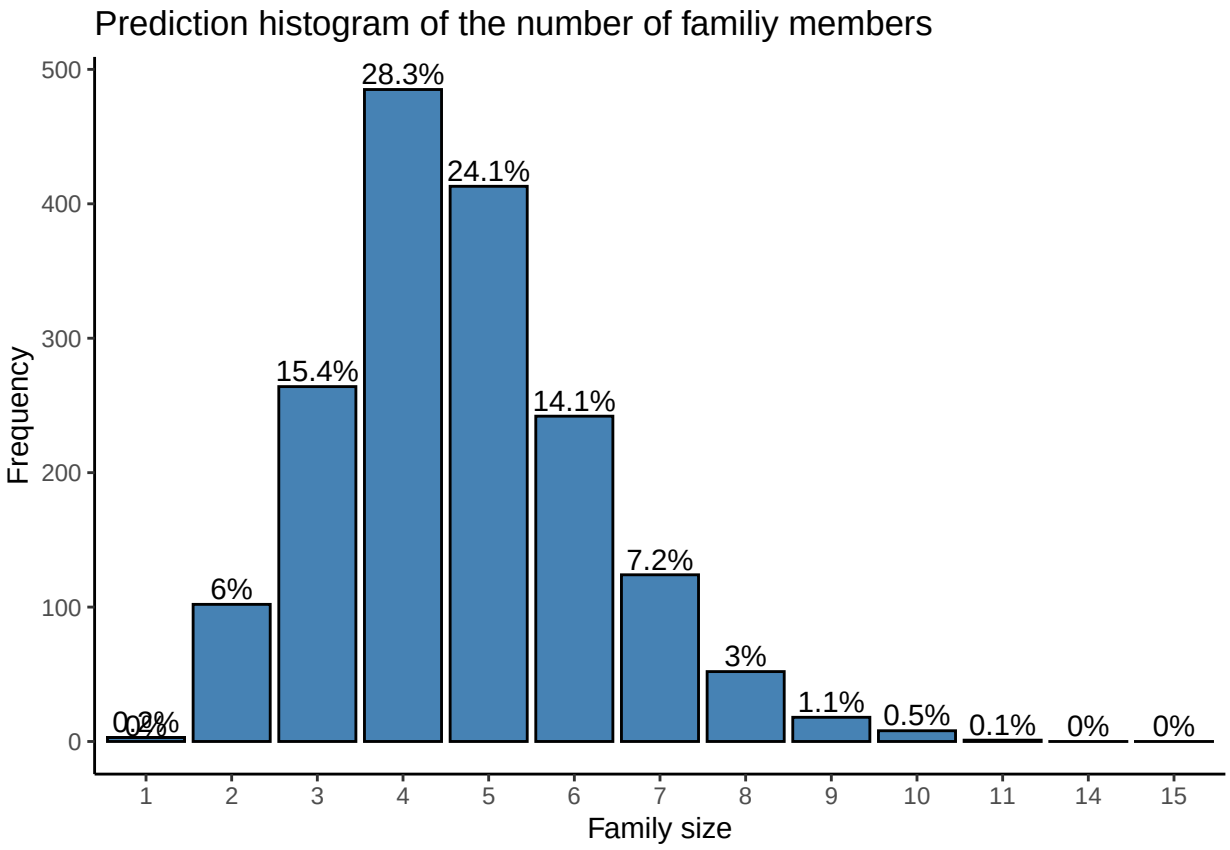


Figure 6: Histogram of predictions of the number of family members

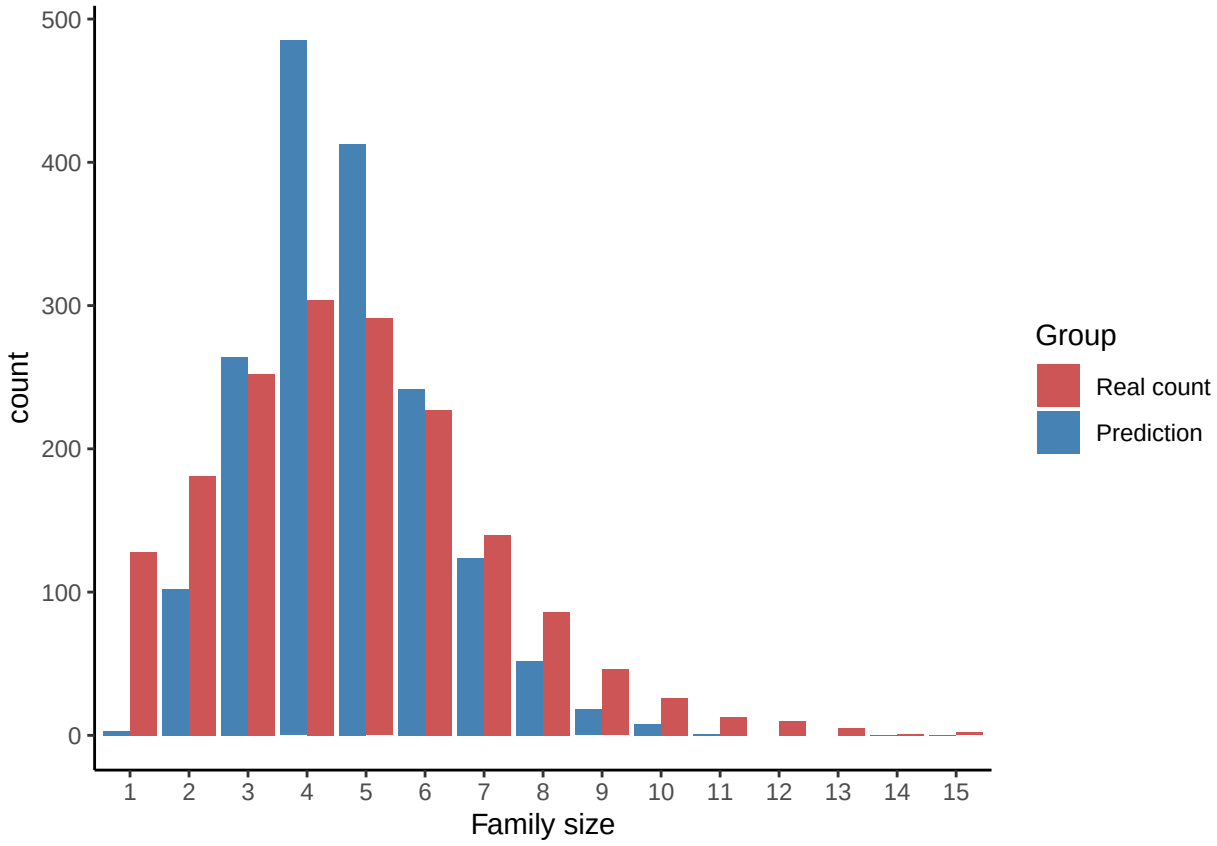


Figure 7: Comparison between real data and predictions

Future Work

We can model the count data using other Possession variations such as Quasi Possession, when there is an over-dispersion (variance is greater than the mean).