MA 615 Refrigerator EDA Project

Shuyi Jiang, Liz Morris, Shan Shan, Luna Tang

10/10/2017

Goal of the project:

Exploring the EIA data, we figured out that climates and regions are the least significant factors since very few homes do not use refrigerator. All of these homes are not from the coldest regions of the US, where people might not need to use refrigerator. Household income influences living condition, including housing unit type, owner/renter status and other factors. In this case, we believe that household income should be the most important factor for planning and marketing refrigerators. We are going to look at the type, size, age and number of refrigerators owned when the population is broken down by income.

Part I. number of refrigerator vs household income vs household income

```
use <- read.csv("use.csv", header = TRUE)
```

```
## Warning in read.table(file = file, header = header, sep = sep, quote =
## quote, : incomplete final line found by readTableHeader on 'use.csv'
```

```
use1 <- read.csv("use1.csv", header = TRUE)
incomelevel <- c("under 20", "20 to 40", "40 to 60", "60 to 80",
                 "80 to 100", "100 to 120", "120 to 140", "140 or more")
use1$income <- factor(use1$income, levels = incomelevel)
```

Five number summary: number of refrigerator vs household income

```
summary(use)
```

```
##  number.of.refrigerator    under.20          X20.to.40         X40.to.60
##  0       :1            Min.   : 0.400   Min.   : 0.000   Min.   : 0.0
##  1       :1            1st Qu.: 1.800   1st Qu.: 2.950   1st Qu.: 2.5
##  2 or more:1           Median : 3.200   Median : 5.900   Median : 5.0
##                        Mean   : 7.467   Mean   : 8.967   Mean   : 6.2
##                        3rd Qu.:11.000   3rd Qu.:13.450   3rd Qu.: 9.3
##                        Max.   :18.800   Max.   :21.000   Max.   :13.6
##     X60.to.80         X80.to.100        X100.to.120       X120.to.140
##  Min.   : 0.00    Min.   :0.000    Min.   :0.000    Min.   :0.00
##  1st Qu.: 2.60    1st Qu.:2.000    1st Qu.:1.700    1st Qu.:1.25
##  Median : 5.20    Median :4.000    Median :3.400    Median :2.50
##  Mean   : 5.10    Mean   :3.233    Mean   :2.733    Mean   :1.80
##  3rd Qu.: 7.65    3rd Qu.:4.850    3rd Qu.:4.100    3rd Qu.:2.70
##  Max.   :10.10    Max.   :5.700    Max.   :4.800    Max.   :2.90
##    X140.or.more
##  Min.   :0.000
##  1st Qu.:2.550
##  Median :5.100
##  Mean   :3.733
```

```
##  3rd Qu.:5.600
##  Max.   :6.100
```

Boxplot: number of refrigerator vs household income

The boxplot indicates that around half of the families with household income of 80000 or more have more than 1 refrigerator. The majority of the richest families with household income of 140000 or more have more than 1 refrigerator.

```r
boxplot(use, ylab="number of refrigerator", xlab="household income")
```
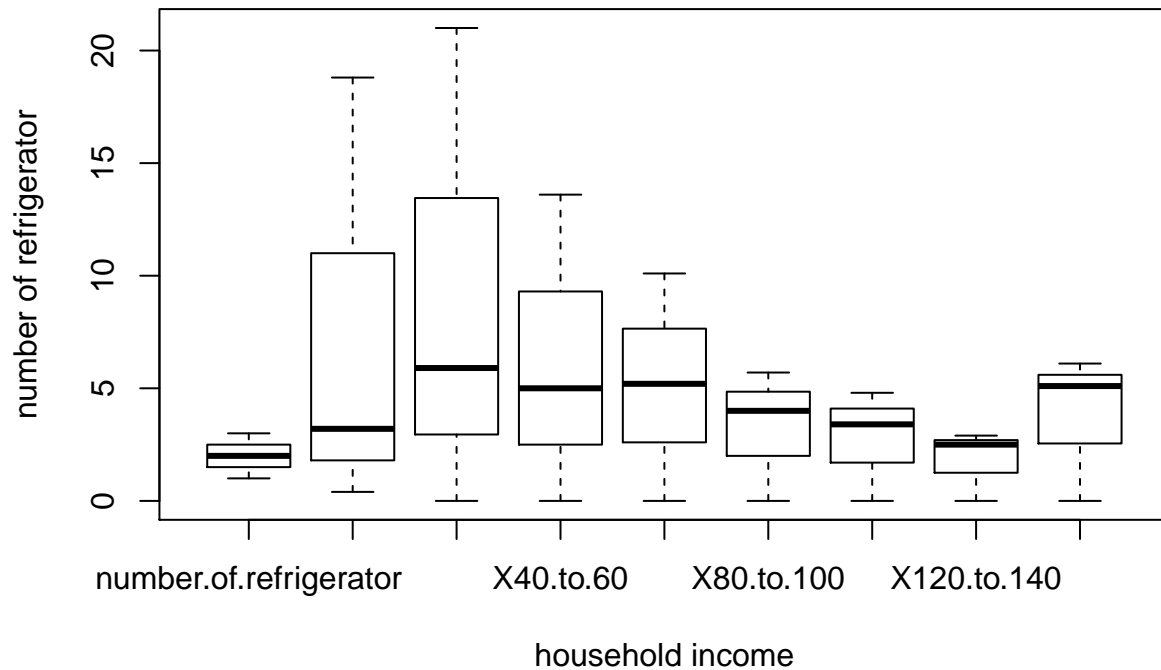

histogram: number of refrigerator vs household income

The histogram indicates the number of refrigerators own by family with different household income. All the families that do not use refrigerator have household income under 20000.

```r
ggplot(data = use1, aes(x= income, y = household, fill= number.of.refrigerator))+
  geom_bar(stat="identity")
```
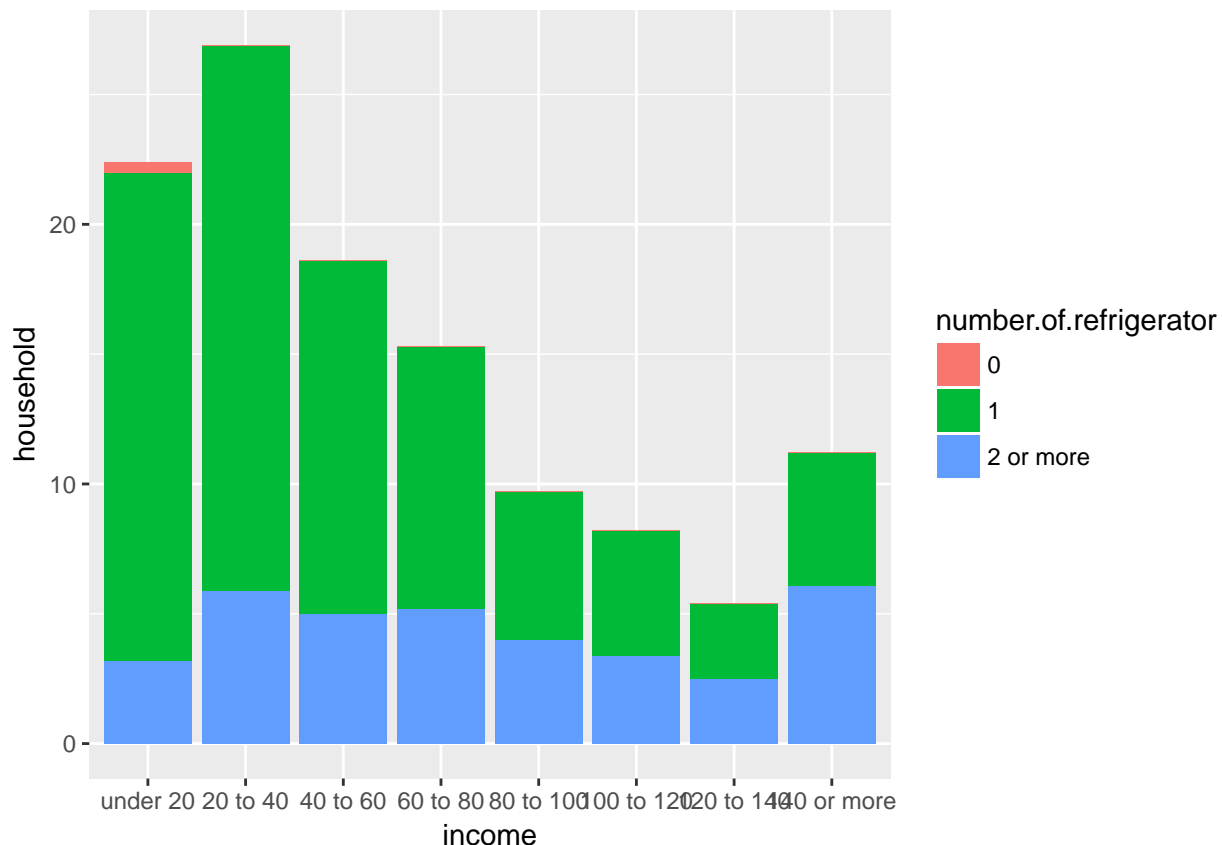
Summary: number of refrigerator owned by population breaking down by household income

Very limited (less than 1%) homes in the US do not have a refrigerators and they come from household with lowest income. The majority of American homes have only 1 refrigerator. As household income increases, the percentage of having more than 1 refrigerator in each income level increases. There is a jump between income levels of 60,000 to 80,000 and 80,000 to 100,000. Nearly half of homes with household income of 80,000 to 140,000 have more than 1 refrigerator. More than half of homes(54.5%) with income of 140,000 or more have at least 2 refrigerators. Thus, the second refrigerator should mainly target at homes with more than 80,000 household income.

Part II. refrigerator size vs household income

Five number summary: number of refrigerator size vs household income

```
size2 <- read.csv("size2.csv", header = TRUE)
summary(size2)
```

```
##                                    X          X..20
##  Do not use a refrigerator        :1   Min.   : 0.000
##  Half-size or compact             :1   1st Qu.: 0.425
##  Large (22.6 to 29.5 cubic feet)  :1   Median : 1.450
##  Medium (17.6 to 22.5 cubic feet) :1   Mean   : 3.717
##  Small (less than 17.6 cubic feet):1   3rd Qu.: 5.475
```

```
## Very large (more than 29.5 cubic feet):1   Max.   :12.500
##     X20.40           X40.60           X60.80           X80.100
## Min.   : 0.000   Min.   : 0.000   Min.   :0.000   Min.   :0.000
## 1st Qu.: 0.225   1st Qu.: 0.050   1st Qu.:0.000   1st Qu.:0.050
## Median : 1.100   Median : 0.550   Median :0.300   Median :0.250
## Mean   : 4.500   Mean   : 3.100   Mean   :2.517   Mean   :1.617
## 3rd Qu.: 7.375   3rd Qu.: 5.475   3rd Qu.:5.175   3rd Qu.:3.225
## Max.   :15.400   Max.   :10.500   Max.   :7.800   Max.   :5.000
##     X100.120         X120.140         X.140
## Min.   :0.000   Min.   :0.00   Min.   :0.000
## 1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.075
## Median :0.100   Median :0.00   Median :0.450
## Mean   :1.333   Mean   :0.85   Mean   :1.867
## 3rd Qu.:2.375   3rd Qu.:1.50   3rd Qu.:2.925
## Max.   :4.700   Max.   :3.10   Max.   :6.600
```

histogram: number of refrigerator vs household income

(Note: we denote the refrigerator as the first refrigerator herewith)

The histogram displays that the number of refrigerator-size owned by family varies according to household income: the higher the income, the larger the size of the refrigerator. Specifically, lower-income households, such as income under 20 thousand dollars, are more likely to "not use a refrigerator"" or use the "small size"" refrigerators; while, middle-income households would obtain the medium size refrigerators; and those with income over 140 thousand dollars prefer "very large size"" refrigerators over other sizes. Medium size refrigerators are purchased by each income groups, but be the most popular one among the lower-income groups like "20-40 thousand" and "40-60 thousand".

```r
size1<-read.csv("size.csv", header=TRUE)
size <-read.csv("size.csv", header=TRUE)

incomeorder<- c('<20', '20-40', '40-60', '60-80','80-100','100-120', '120-140','>140')
sizeorder<-c('Half-size or compact','Small (less than 17.6 cubic feet)',
            'Medium (17.6 to 22.5 cubic feet)','Large (22.6 to 29.5 cubic feet)',
            'Very large (more than 29.5 cubic feet)','Do not use a refrigerator')
size1$size<- factor(size1$size, levels=sizeorder)
size1$income <- factor(size1$income, levels=incomeorder)
ggplot (data=size1, aes(x=income,y=unit,fill=size))+geom_bar(stat="identity")
```
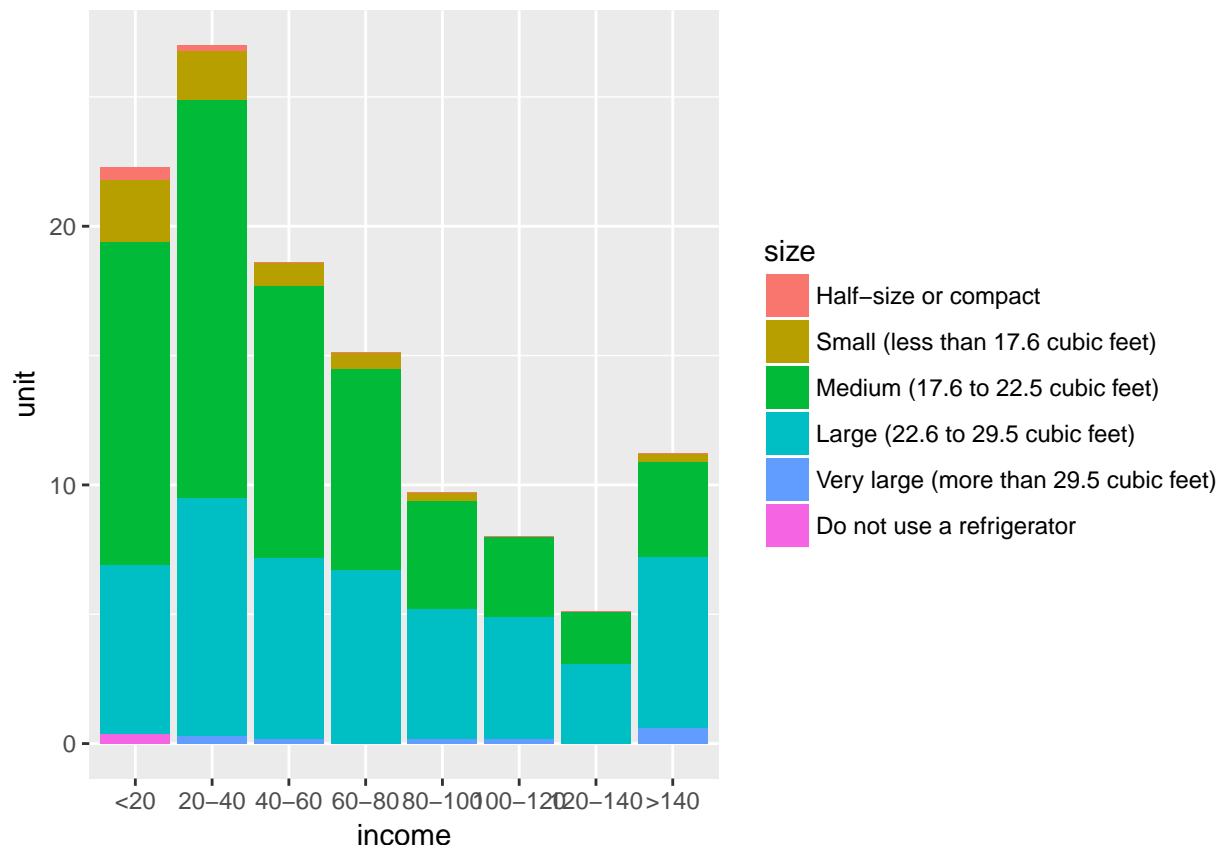
Summary: household income influnces the selection of refrigerator-size positively; and thus we recommend the income-oriented marketing strategy with regards to the refrigerator-size.

The study of the relationship between refrigerator-size and household income shows that different income groups have the different preferences for refrigerator-size; and the higher the income, the larger size of refrigerators. Thus, we could enhance the promotion of large size refrigerator among those high-income consumers; and also further our strategy of medium-size refrigerators among the lower-income groups. Additionally, since "very large refrigerators"" presents a promising response among the "annual income more than 140 thousand" group, we could further its market share correspondently.

Part III. refrigerator types vs household income

Five number summary: refrigerator types vs household income

```
type <- read.csv("type.csv")
summary(type)
```

```
##                             X      Less.than..20.000 X.20.000.to..39.999
##  Do not use a refrigerator:1   Min.   : 0.00     Min.   : 0.000
##  Half-size/compact        :1   1st Qu.: 0.45     1st Qu.: 0.250
##  One door                 :1   Median : 1.10     Median : 2.200
##  Three or more doors      :1   Mean   : 3.20     Mean   : 3.857
##  Two doors, bottom freezer:1   3rd Qu.: 3.35     3rd Qu.: 5.250
```

5

```
## Two doors, side by side  :1   Max.   :13.70     Max.   :13.800
## Two doors, top freezer    :1
## X.40.000.to..59.999 X.60.000.to..79.999 X.80.000.to..99.000
## Min.   :0.000        Min.   :0.000        Min.    :0.000
## 1st Qu.:0.150        1st Qu.:0.150        1st Qu.:0.200
## Median :1.200        Median :1.300        Median :0.500
## Mean   :2.629        Mean   :2.186        Mean    :1.386
## 3rd Qu.:3.950        3rd Qu.:3.850        3rd Qu.:2.300
## Max.   :9.000        Max.   :6.000        Max.    :4.200
##
## X.100.000.to..119.999 X.120.000.to...139.999 X.140.000.or.more
## Min.   :0.000         Min.   :0.0000         Min.    :0.000
## 1st Qu.:0.200         1st Qu.:0.1000         1st Qu.:0.100
## Median :0.500         Median :0.2000         Median :0.800
## Mean   :1.157         Mean   :0.7429         Mean    :1.586
## 3rd Qu.:2.100         3rd Qu.:1.3000         3rd Qu.:2.600
## Max.   :3.000         Max.   :2.2000         Max.    :4.900
##
```

Boxplot: refrigerator types vs household income

The boxplot indicates that the number of different types of refrigerator that a household owns varies as household income varies. Rich households which have larger than $80,000 income has less types of refrigerators that those with median income($20,000-$80,000). Families with less than 20,000 has the most types of refrigerators.

```r
boxplot(type,xlab="income",ylab="types")
```
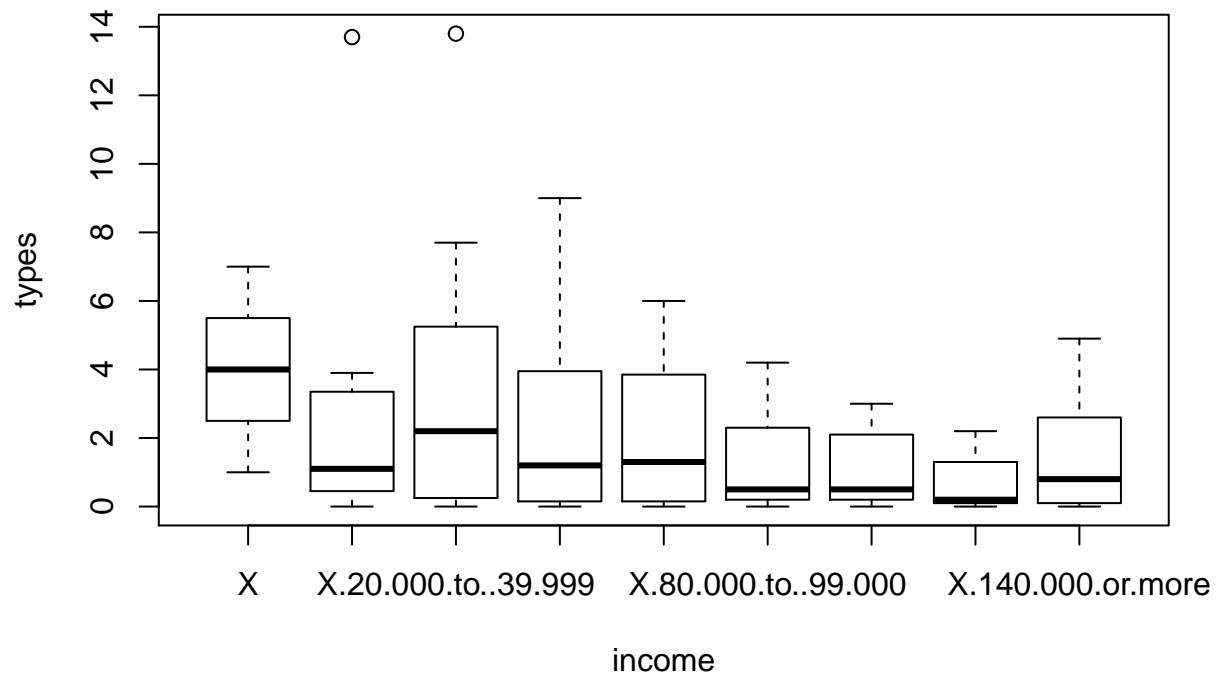
histogram: refrigerator types vs household income

The histogram indicates the types of refrigerators own by families with different household income. According to the graph, two doors, top refrigerator is the most used types of families with less than 20,000 income, 20,000-40,000 income, 40,000-60,000 income, and 60,000-80,000 income. Other households use two doors, side by side refrigerators the most. Households with less than 20,000 income and households with 20,000-40,000 income has the most unites of refrigerators.

```
type1 <- read.csv("type1.csv")
ggplot(data=type1, aes(x=income, y=unites, fill=type)) + geom_bar(stat="identity")
```
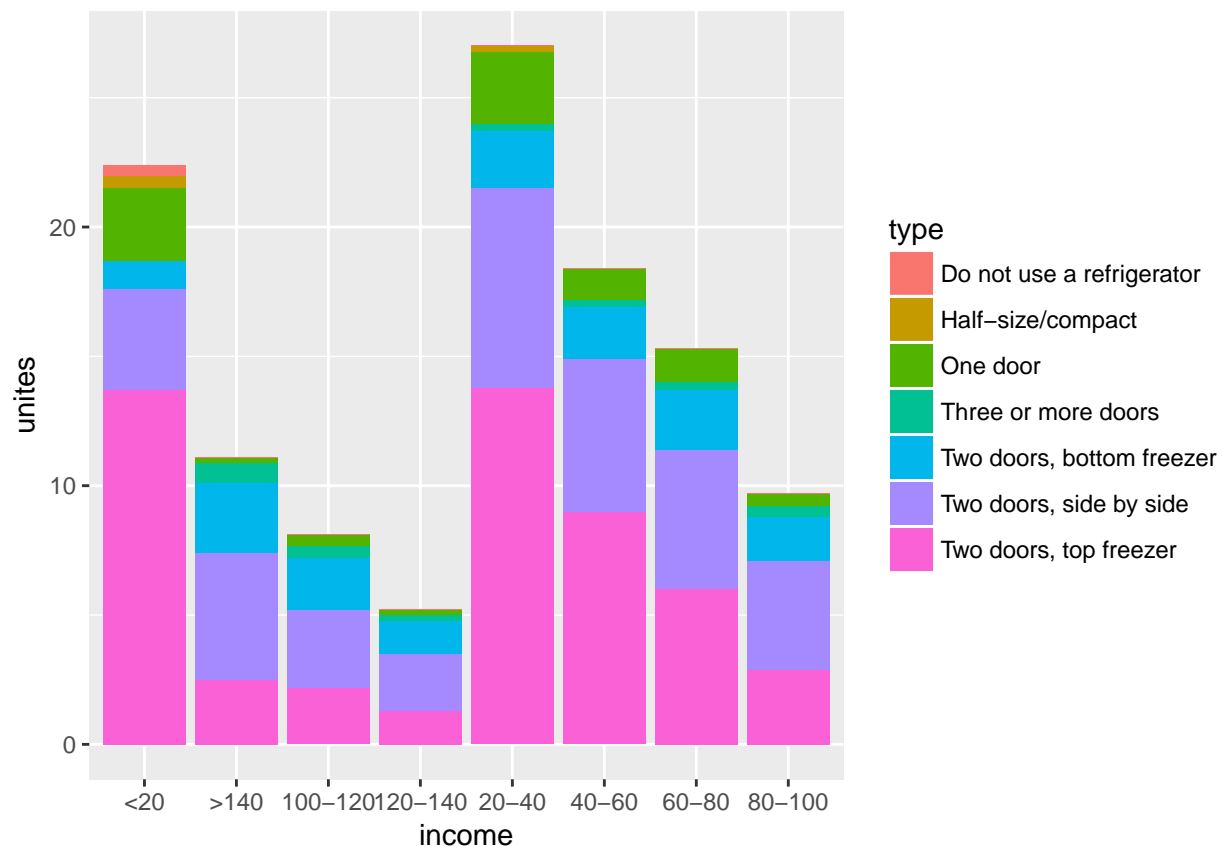

Summary: types of refrigerator owned by families breaking down by household income

Two doors, top refrigerator and two doors, side by side refrigerator are the most commonly used refrigerator types by majority of all households.Thus, marketing should be focusing on two doors, top refrigerator and two doors, side by side refrigerator. households of less than 20,000 income and households with 20,000-40,000 (median income families) have the most number of refrigerators, which indicates marketing should be focusing on these two groups of households.Half-size/compact regrigerators are only used by low income families. Companies wishing to sell half-size/compact refrigerators may focus on low income families by reducing the price of the products.

Part IV: Age of refrigerator vs household income

This information is useful because it shows how often people are buying refrigerators and who may be in the market for a new one.

Proportion Summary

```r
age1 <- read.csv("rg age1.csv", header = TRUE)
age <- read.csv("rg age.csv", header = TRUE)
colnames(age) <- c('Age','Under 20', '20 to 40', '40 to 60', '60 to 80',
                   '80 to 100', '100 to 120', '120 to 140', '140+')

total <- sum(age1$Units)
#aggregate(age1$Units, by=list(Category=age1$Age), FUN=sum)
age1 %>% group_by(Age) %>% summarise(proportion = sum(Units) / total)
```

```
## # A tibble: 7 x 2
##                          Age  proportion
##                       <fctr>       <dbl>
## 1       10 to 14 years old 0.188936170
## 2       15 to 19 years old 0.063829787
## 3         2 to 4 years old 0.222978723
## 4     20 or more years old 0.039148936
## 5          5 to 9 years old 0.348085106
## 6 Do not use a refrigerator 0.003404255
## 7    Less than 2 years old 0.133617021
```

This summary shows the proportions of refrigerator ages, regardless of income level. This is useful as it shows how old most people's refrigerators are.

```r
incomeorder <- c('Under 20', '20 to 40', '40 to 60', '60 to 80',
                 '80 to 100', '100 to 120', '120 to 140', '140+')
ageorder <- c('Less than 2 years old', '2 to 4 years old', '5 to 9 years old',
              '10 to 14 years old', '15 to 19 years old', '20 or more years old',
              'Do not use a refrigerator')

age1$Age <- factor(age1$Age, levels = ageorder)
age1$Income <- factor(age1$Income, levels=incomeorder)
```

Boxplot

```r
boxplot(age[, 2:8], xlab = "Income bracket (in thousands of dollars)",
        ylab = "Units (in thousands)")
```
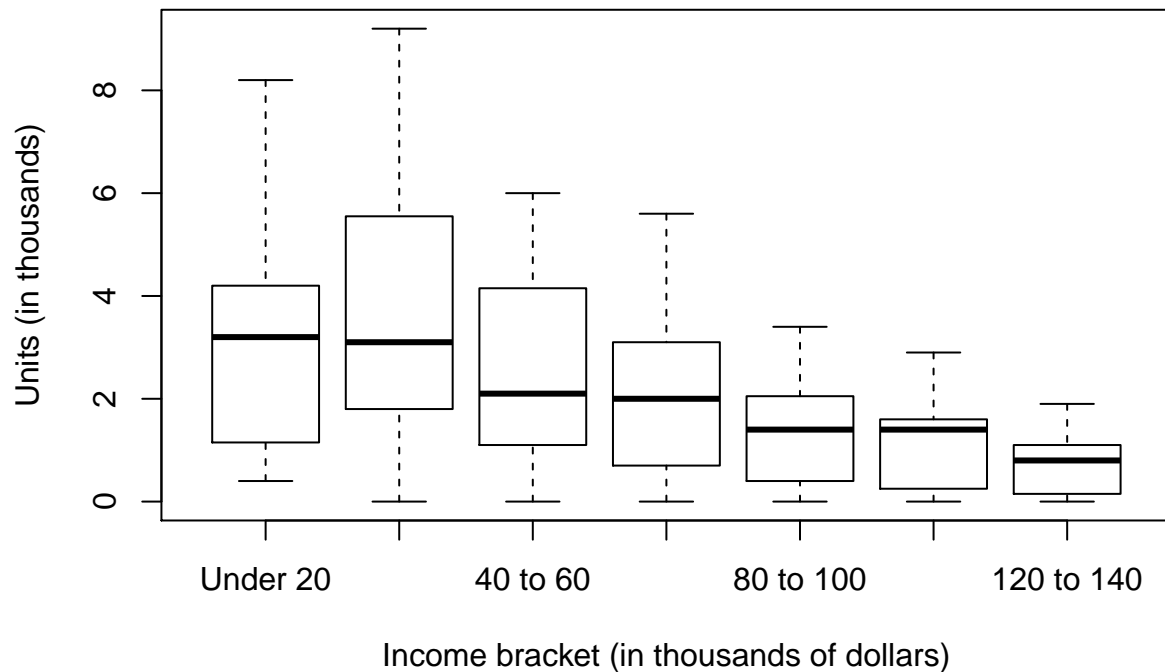
While this boxplot does not give much information about the ages of refrigerators, it does show which income groups have higher variations of refrigerator ages because each data point is the number of units of refrigerators in an age category.

Bargraph

```
ggplot(data=age1, aes(x=Income, y=Units, fill=Age)) +
  geom_bar(stat="identity") +
  labs(x= "Income (in thousands of dollars)", y = "Units (in thousands)")
```
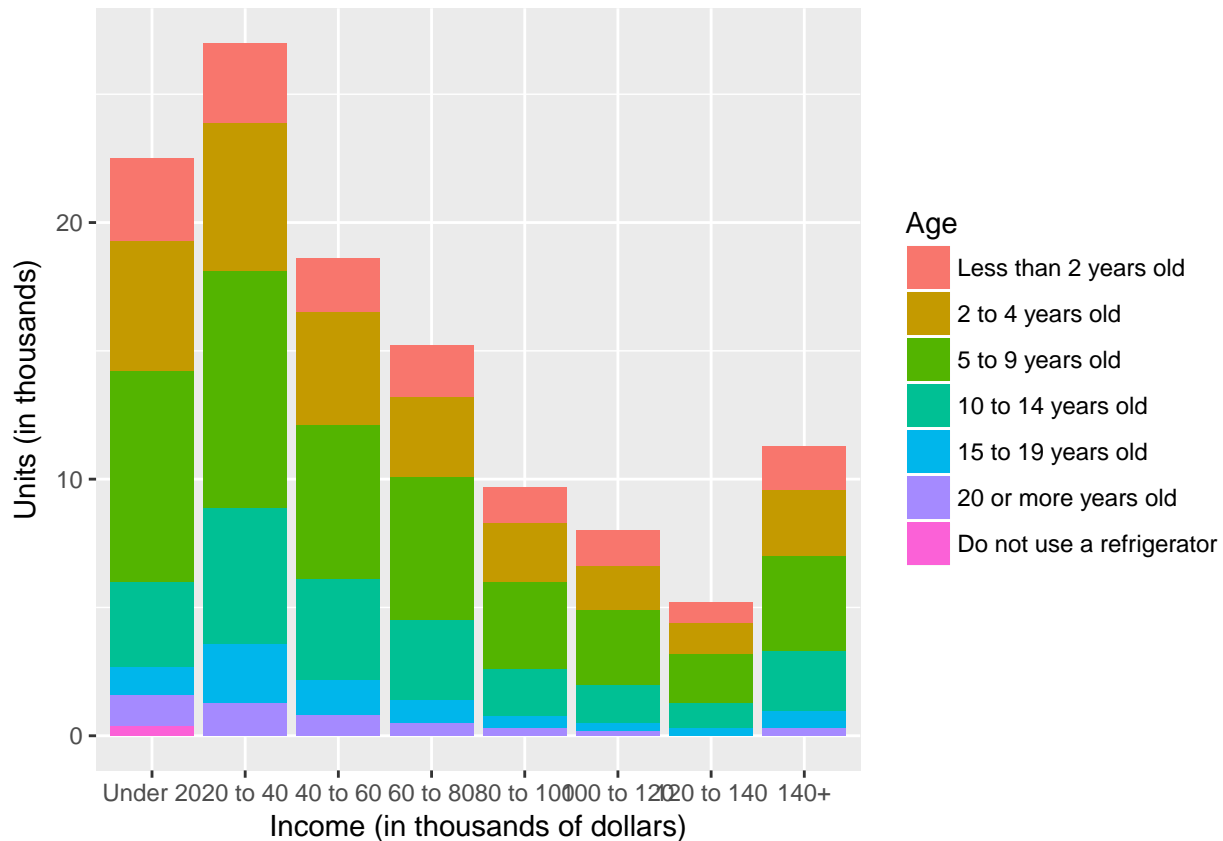
Looking at this plot, you can see that the largest number of people with new refrigerators as well as with very old refrigerators are in the lower income brackets. People with higher income have refrigerators that are between 5 and 15 years old.

Summary

Overall, the highest proportion of people have refrigerators that are between 5 and 9 years old, suggesting that most people replace their refrigerator every 10 or so years. Additionally, we can see that the income brackets with the oldest refrigerators are the lowest ones, meaning that they may be in the market for a new refrigerator. The lowest income brackets also have higher proportions of people who have newer refrigerators, which may indicate that they buy refrigerators more often because they cannot afford to buy high quality refrigerators. Overall, it would seem that those in the lowest income brackets are most likely to be in the market for a new refrigerator and thus should be a main target of any marketing campaigns.

Collaboration

We explored the EIA data online and shared our observations with team members. We disccued what we were interested in. And then we decided the variables we were going to look at, the method to use and the graph to interpret. Since we have 4 variables, each of us did the coding, graphing and interpreting for 1 of them. We combined our script together and finished the summary and formating together.