

615 project

Yaqi Huang, Ying li, Liuqian Zhang

2017/10/6

Our team has been asked to explore the data related to the refrigerator in order to provide strategy for planning and marketing refrigerator. The team of us would look up three important factors such as, if households owning a fridge or not, the size of the fridge and the age of the most-used fridge, which would indicate the willingness of households on purchasing refrigerator.

The question we posted is “How the housing units distributed under these three category (fridge, size, age) and what we could suggest for new strategy for marketing refrigerator?”

The main method we used is EDA, and produced stacked bar charts to see the distribution under each factor.

Climate Region

Data import and cleaning

```
library(ggplot2)
library(tibble)
library(readr)
library(reshape2)

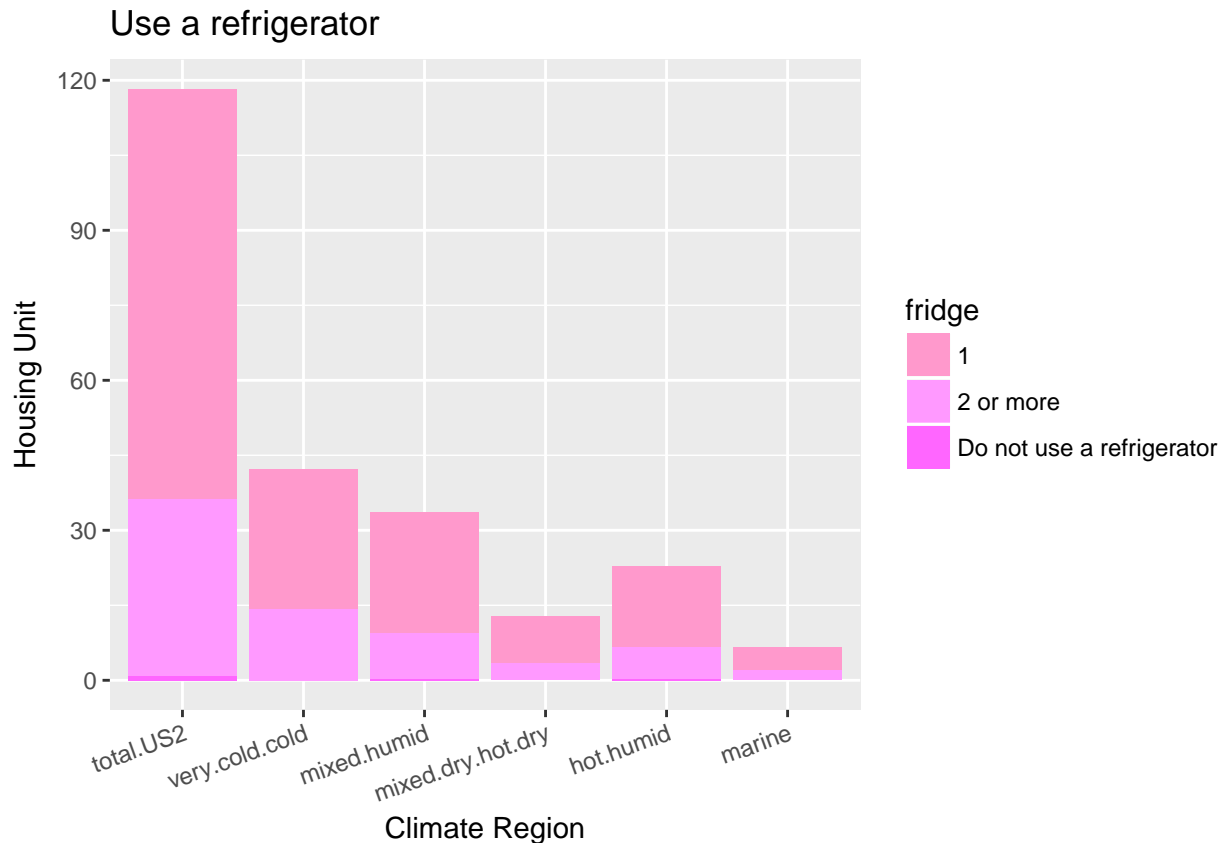
f11<-read.csv("f11.csv",header = TRUE,na=c("Q","N"))
f11.m <- melt(f11,id.vars = "fridge")

f11.m[, 1] <- sapply(f11.m[, 1], as.factor)
f11.m[, 2] <- sapply(f11.m[, 2], as.factor)
f11.m[, 3] <- sapply(f11.m[, 3], as.numeric)
```

Owning a fridge or not

```
gg1 <- ggplot() + geom_bar(aes(y = value, x = variable, fill = fridge), data = f11.m, stat = "identity",
  theme(axis.text.x = element_text(angle = 20,hjust=1))+
  scale_fill_manual(values = c("#FF99CC","#FF99FF","#FF66FF")) +
  labs(title="Use a refrigerator",x="Climate Region",y="Housing Unit")

gg1
```



From the chart we could tell that the majority of people are currently owning a fridge, and most of them are located in cold areas. Possible explanation for this could be that the population density in cold regions are larger, to have deeper analysis it is necessary to look up the population distribution.

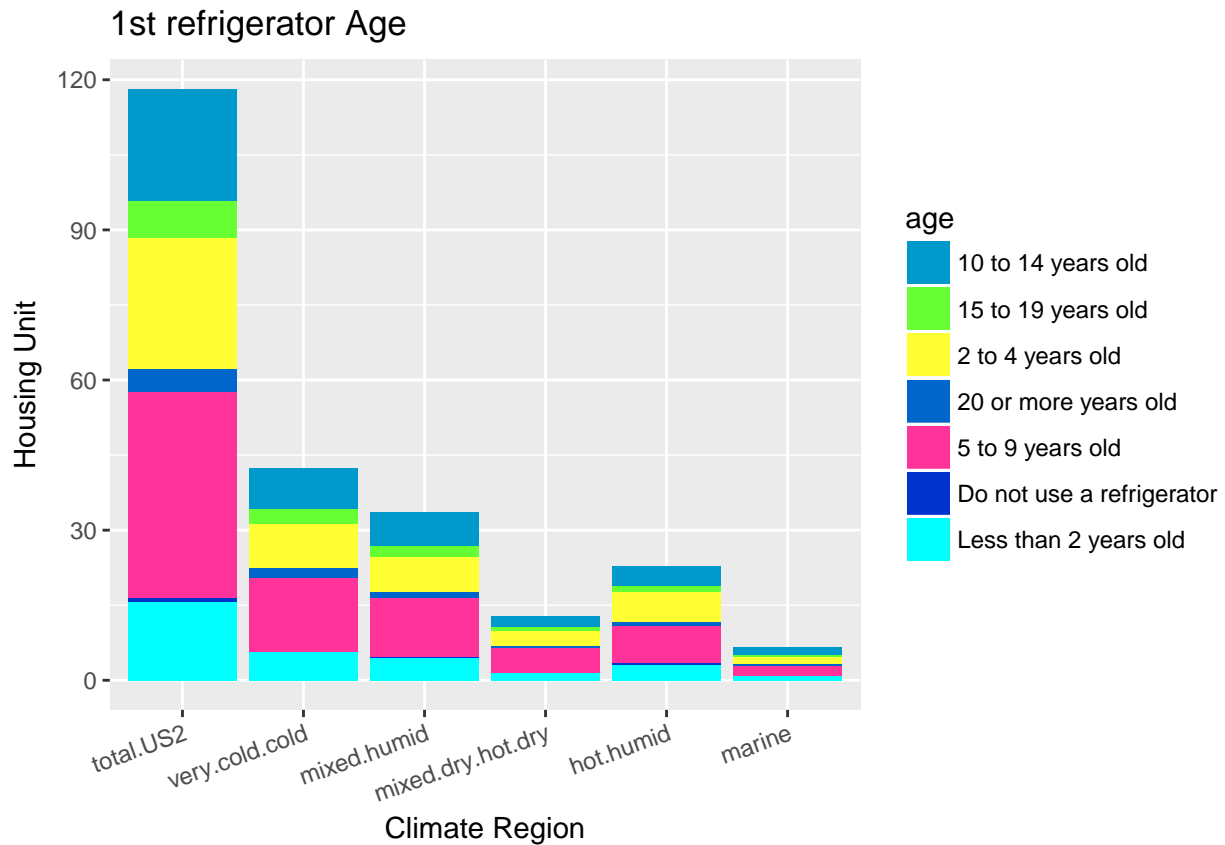
1st fridge age

```
knitr::opts_chunk$set(echo = TRUE)
f11age<-read.csv("f11age.csv",header = TRUE,na=c("Q","N"))
f11.m.age <- melt(f11age,id.vars = "age")

f11.m.age[ , 1] <- sapply(f11.m.age[ , 1], as.factor)
f11.m.age[ , 2] <- sapply(f11.m.age[ , 2], as.factor)
f11.m.age[ , 3] <- sapply(f11.m.age[ , 3], as.numeric)

gg2 <- ggplot() + geom_bar(aes(y = value, x = variable, fill = age), data = f11.m.age, stat = "identity",
  theme(axis.text.x = element_text(angle = 20,hjust=1))+
  scale_fill_manual(values = c("#0099CC","#66FF33","#FFFF33","#0066CC","#FF3399","#0033CC","#00FFFF") ) +
  labs(title="1st refrigerator Age",x="Climate Region",y="Housing Unit")

gg2
```



The chart shows that most of the people could use their fridges between 5 to 9 years, however there is still a large proportion of population only use fridge up to 4 years.

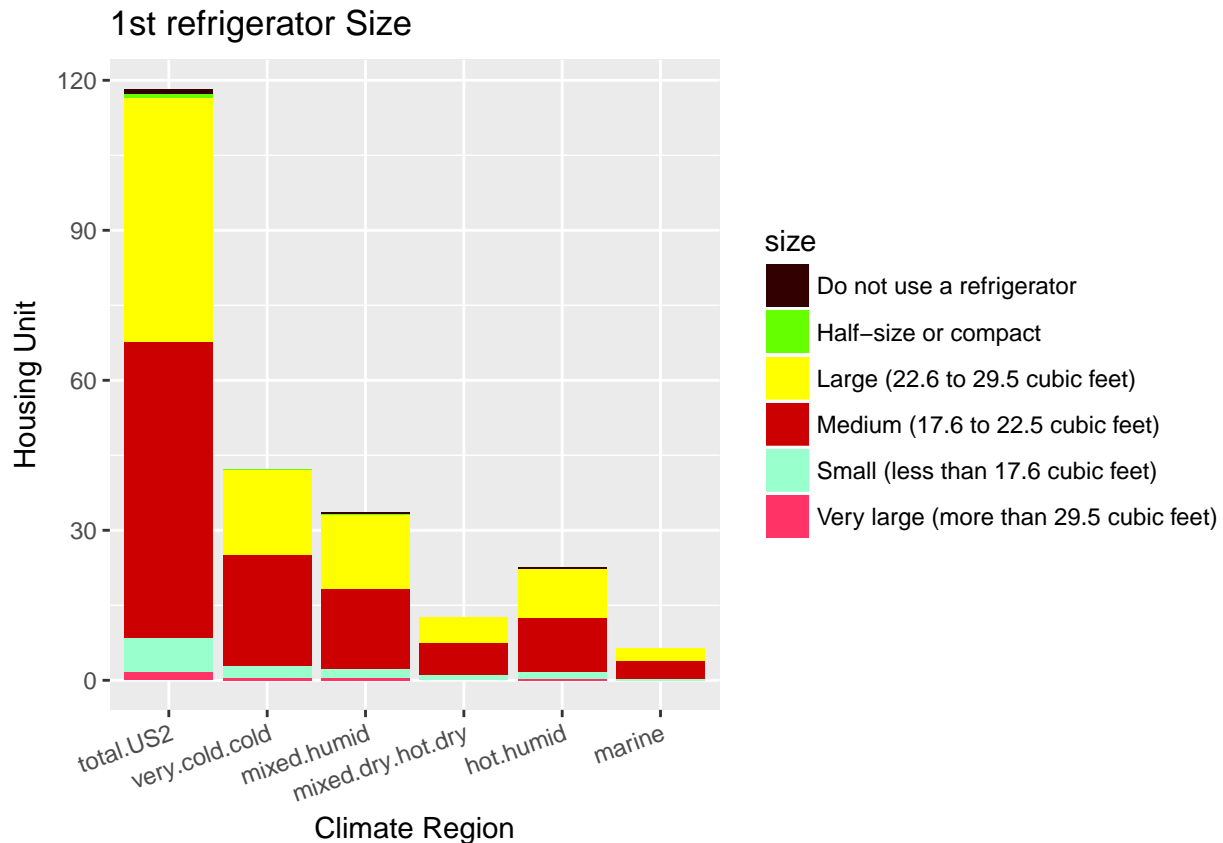
1st fridge size

```
f11size<-read.csv("f11size.csv",header = TRUE,na=c("Q","N"))
f11.m.size <- melt(f11size,id.vars = "size")

f11.m.size[ , 1] <- sapply(f11.m.size[ , 1], as.factor)
f11.m.size[ , 2] <- sapply(f11.m.size[ , 2], as.factor)
f11.m.size[ , 3] <- sapply(f11.m.size[ , 3], as.numeric)

gg3 <- ggplot() + geom_bar(aes(y = value, x = variable, fill = size), data = f11.m.size, stat = "identity")
  theme(axis.text.x = element_text(angle = 20,hjust=1))+
  scale_fill_manual(values = c("#330000","#66FF00","#FFFF00","#CC0000","#99FFCC","#FF3366","#FF3399") ) +
  labs(title="1st refrigerator Size",x="Climate Region",y="Housing Unit")

gg3
```



It seems that the medium-sized fridge is the most owned size for the households, however there is also a large proportion of housing units owning a large size fridge. To analyze this more, it is necessary to look up the number of members in households, which has shown below.

Number of household members

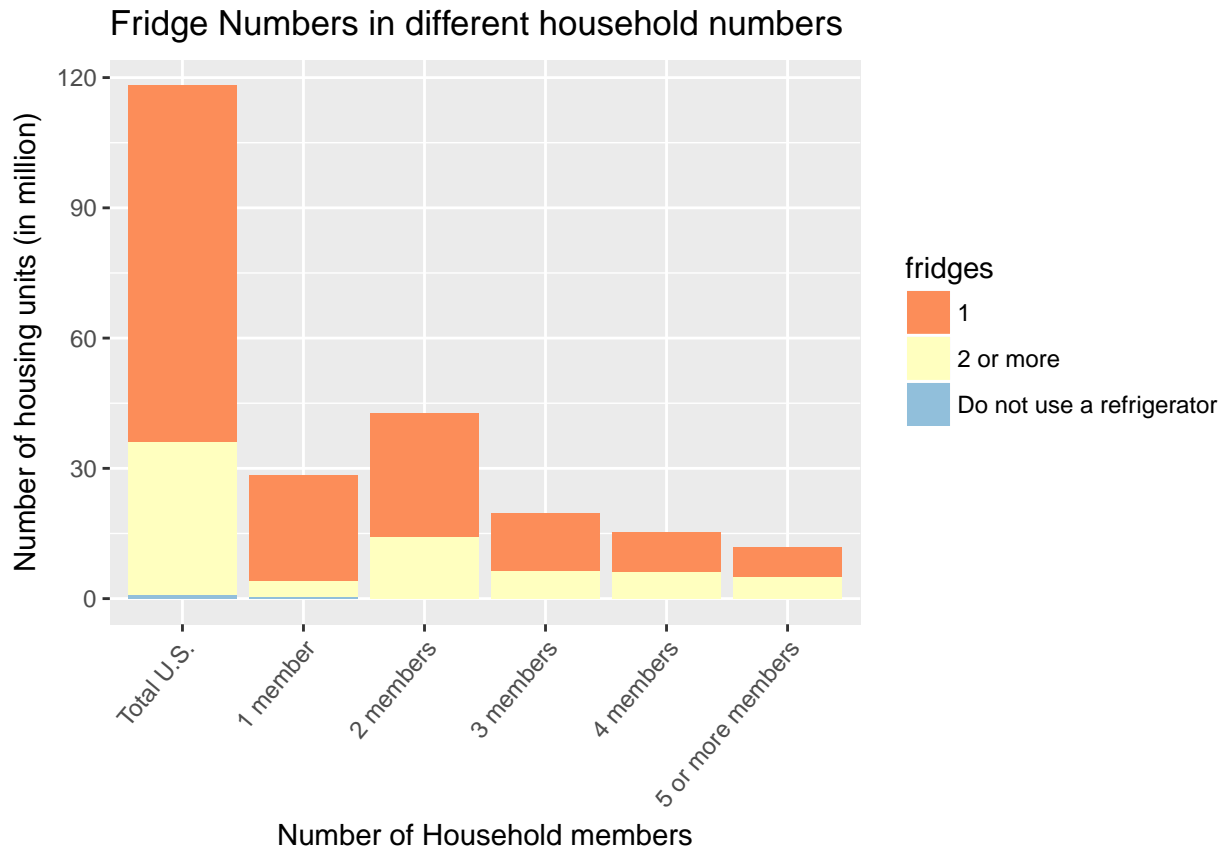
data import and cleaning

```
fridge.numbers <- read_csv("fridge number.csv",na=c("Q","N"))
fridge.size <- read_csv("fridge size.csv",na=c("Q","N"))
fridge.age <- read_csv("fridge age.csv",na=c("Q","N"))

fridge.numbers.m <- melt(fridge.numbers,id.vars = "fridges")
fridge.size.m <- melt(fridge.size,id.vars = "size")
fridge.age.m <- melt(fridge.age,id.vars = "age")
```

fridge numbers stacked bar chart

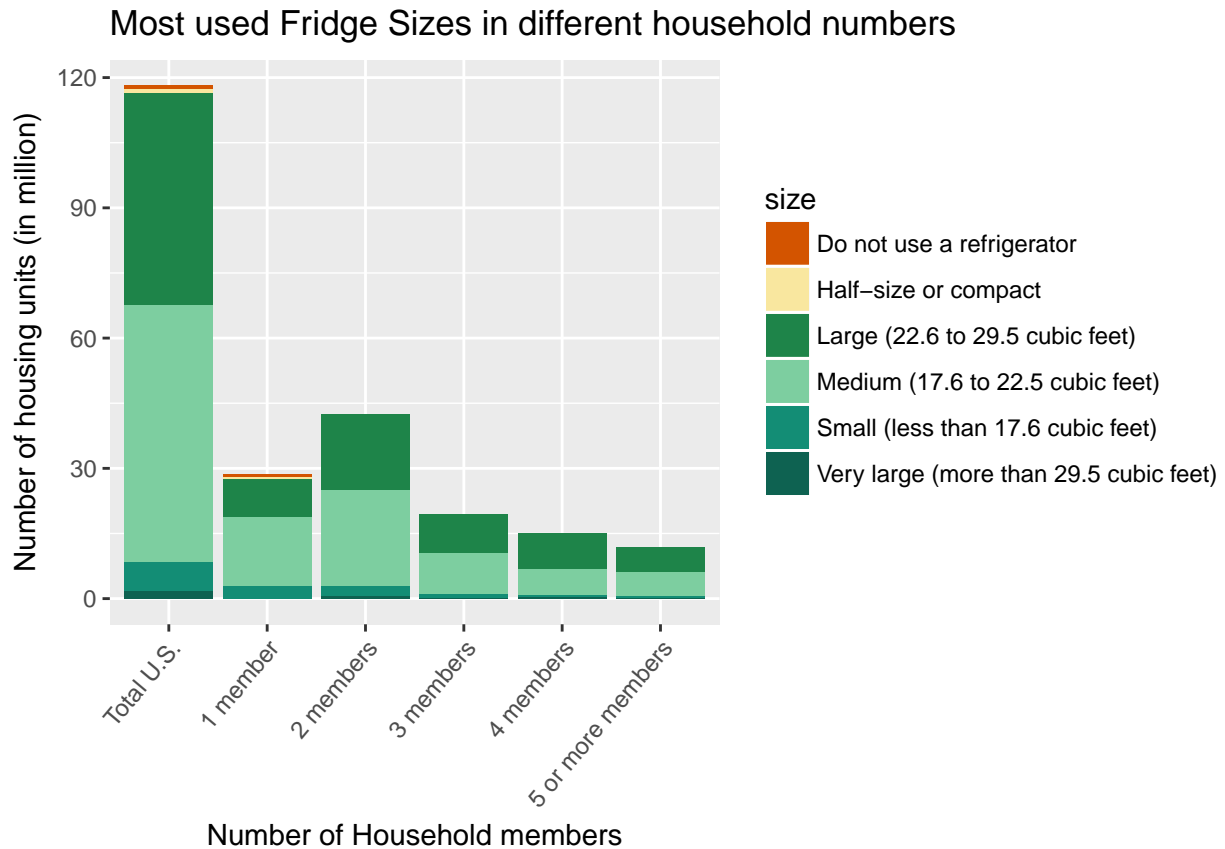
```
ggplot(fridge.numbers.m, aes(x = variable, y = value, fill=fridges)) +
  geom_bar(stat='identity') +
  scale_fill_manual(values = c('#fc8d59', '#ffffbf','#91b1db')) +
  labs(title="Fridge Numbers in different household numbers", x="Number of Household members", y="Number of Fridge")
  theme(axis.text.x = element_text(angle = 50,hjust=1))
```



In U.S., most household have one fridge, some have two or more, and only a few has no fridges. One fridge happens most in two-members households. Two or more fridges happen most in two-members household too. And no fridge situation happens most in one-member households.

fridge size stacked bar chart

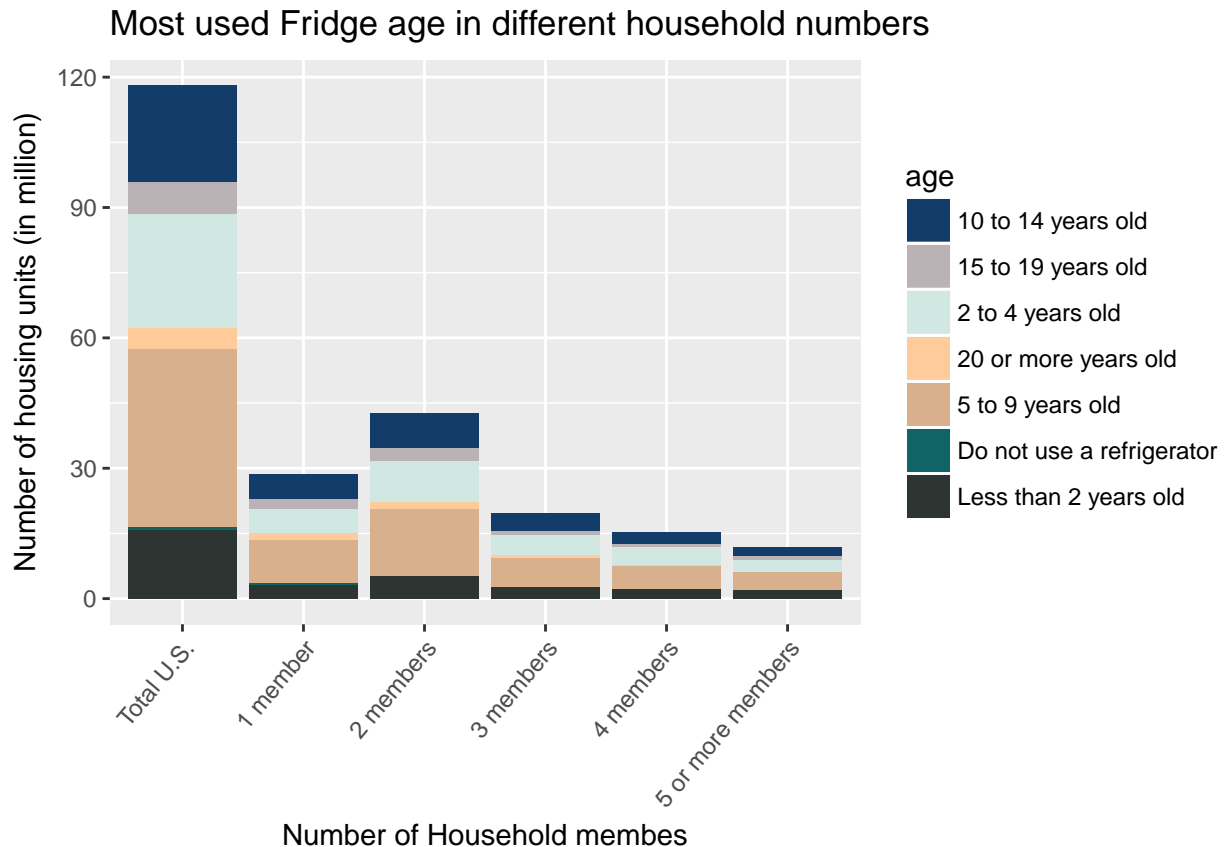
```
ggplot(fridge.size.m, aes(x = variable, y = value, fill=size)) +
  geom_bar(stat='identity') +
  scale_fill_manual(values = c('#D35400', '#F9E79F', '#1E8449', '#7DCEA0', '#138D75', '#0E6251')) +
  labs(title="Most used Fridge Sizes in different household numbers", x="Number of Household members")
  theme(axis.text.x = element_text(angle = 50,hjust=1))
```



In U.S. the most housing units are using medium sized fridges as their most used fridges. Households with 2 members have the most numbers to use medium sized fridges.

fridge ages stacked bar chart

```
ggplot(fridge.age.m, aes(x = variable, y = value, fill=age)) +
  geom_bar(stat='identity') +
  scale_fill_manual(values = c('#123c69', '#bab2b5', '#d1e8e2', '#ffcb9a', '#d9b08c', '#116466', '#2c3531')) +
  labs(title="Most used Fridge age in different household numbers", x="Number of Household membes", y="Number of housing units (in million)") +
  theme(axis.text.x = element_text(angle = 50, hjust=1))
```



In U.S. the most number of housing units have their most used fridge to be 5 to 9 years old. And the least is 20 or more years old. Among them, households with 2 members have the most fridge aged 5 to 9 years old. The least is 20 or more years old for 2 member families.

income and fridge age and size

```
##library
library(ggplot2)
library(reshape2)
```

data import

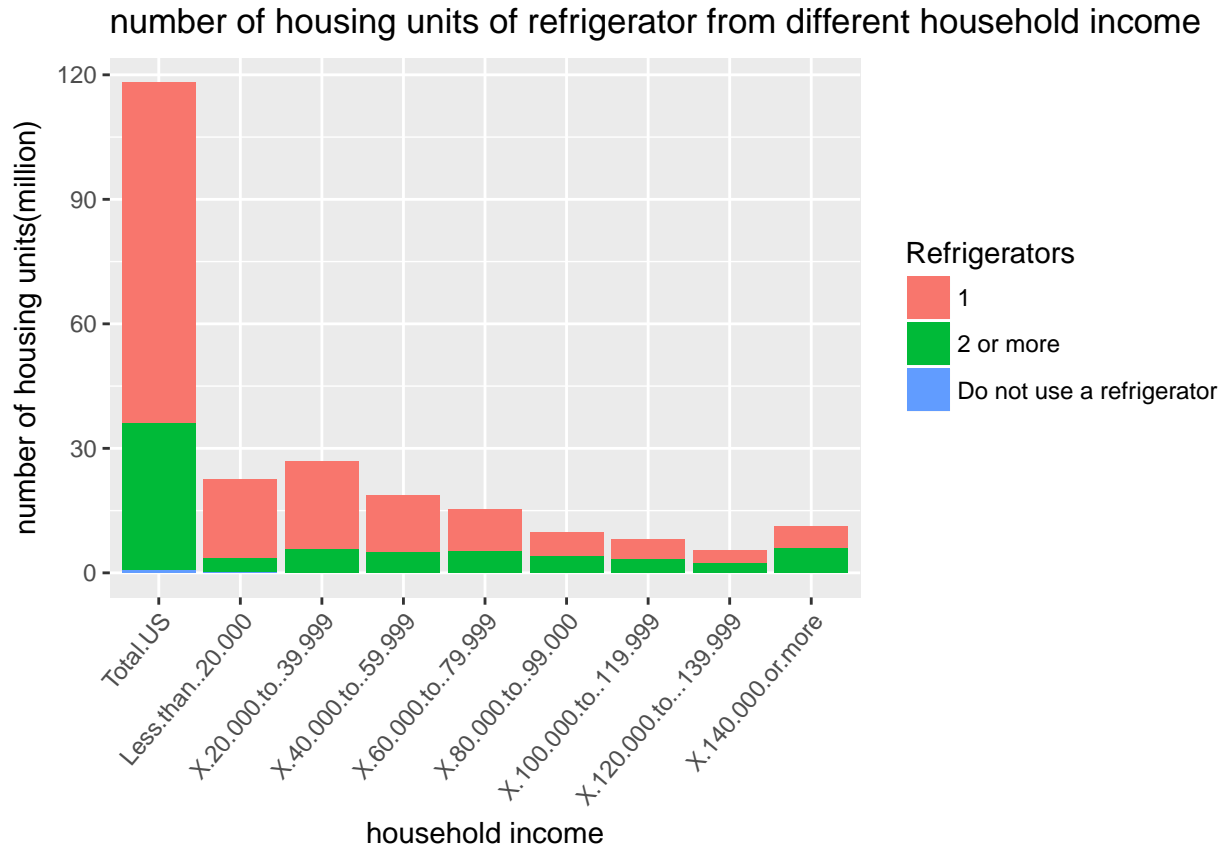
```
fr<-read.csv("New Refrigerators.csv",header = TRUE,na=c("Q","N"))
frsize<-read.csv("New Most-used refrigerator size.csv",header = TRUE,na=c("Q","N"))
frage<-read.csv("New Most-used refrigerator age.csv",header = TRUE,na=c("Q","N"))
```

data cleaning

```
fri<-melt(fr,id.vars = "Refrigerators")
frsize<-melt(frsize,id.vars = "Most.used.refrigerator.size")
friage<-melt(frage,id.vars = "Most.used.refrigerator.age")
```

EDA of fridge

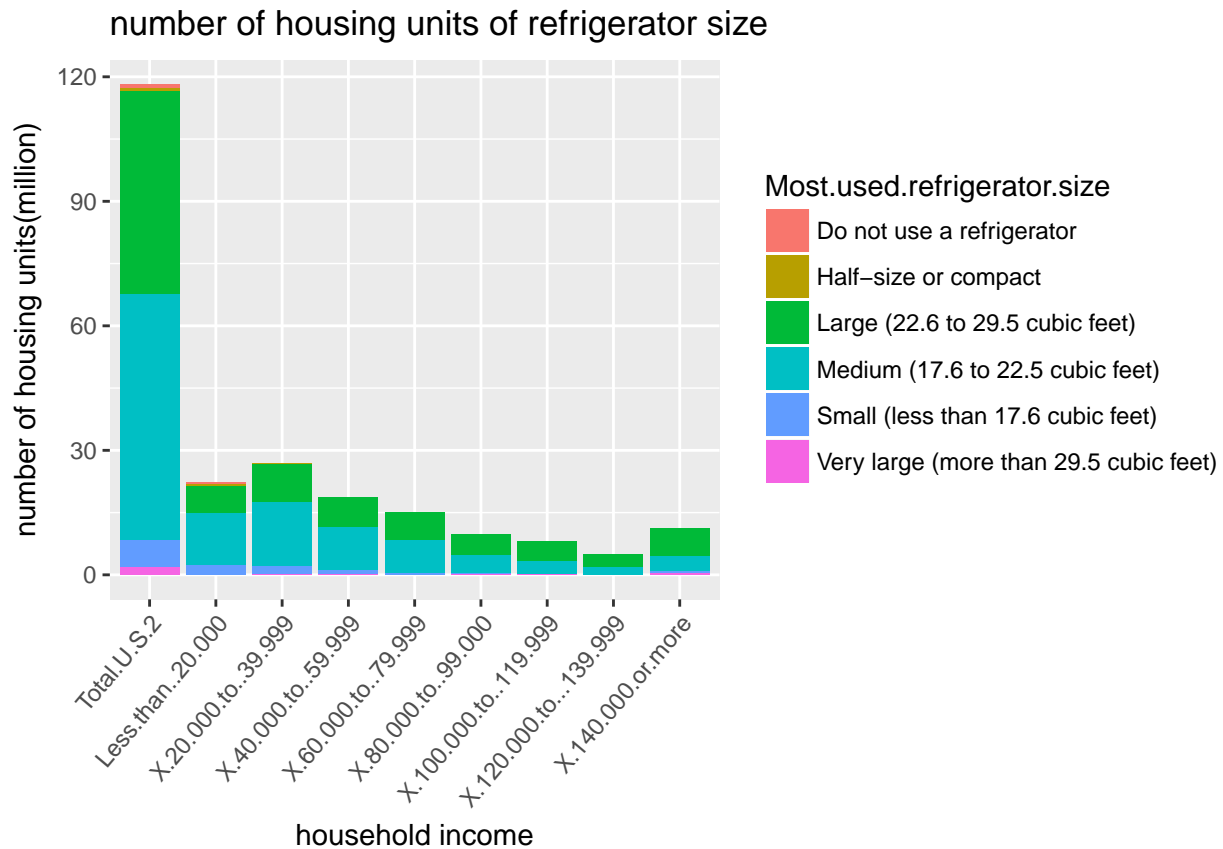
```
fri[, 1] <- sapply(fri[, 1], as.factor)
fri[, 2] <- sapply(fri[, 2], as.factor)
fri[, 3] <- sapply(fri[, 3], as.numeric)
ggfri <- ggplot() + geom_bar(aes(y = value, x = variable, fill = Refrigerators), data = fri, stat = "identity")
ggfri
```



From the first histogram we created, in U.S., most of households that have the different income normally have one fridge, and some have 2 or more fridge, based on the histogram nearly no household do not use a fridge. Moreover, the households that have the income of 20000 to 39999 have the highest household numbers that have on or more fridges.

EDA of fridge size

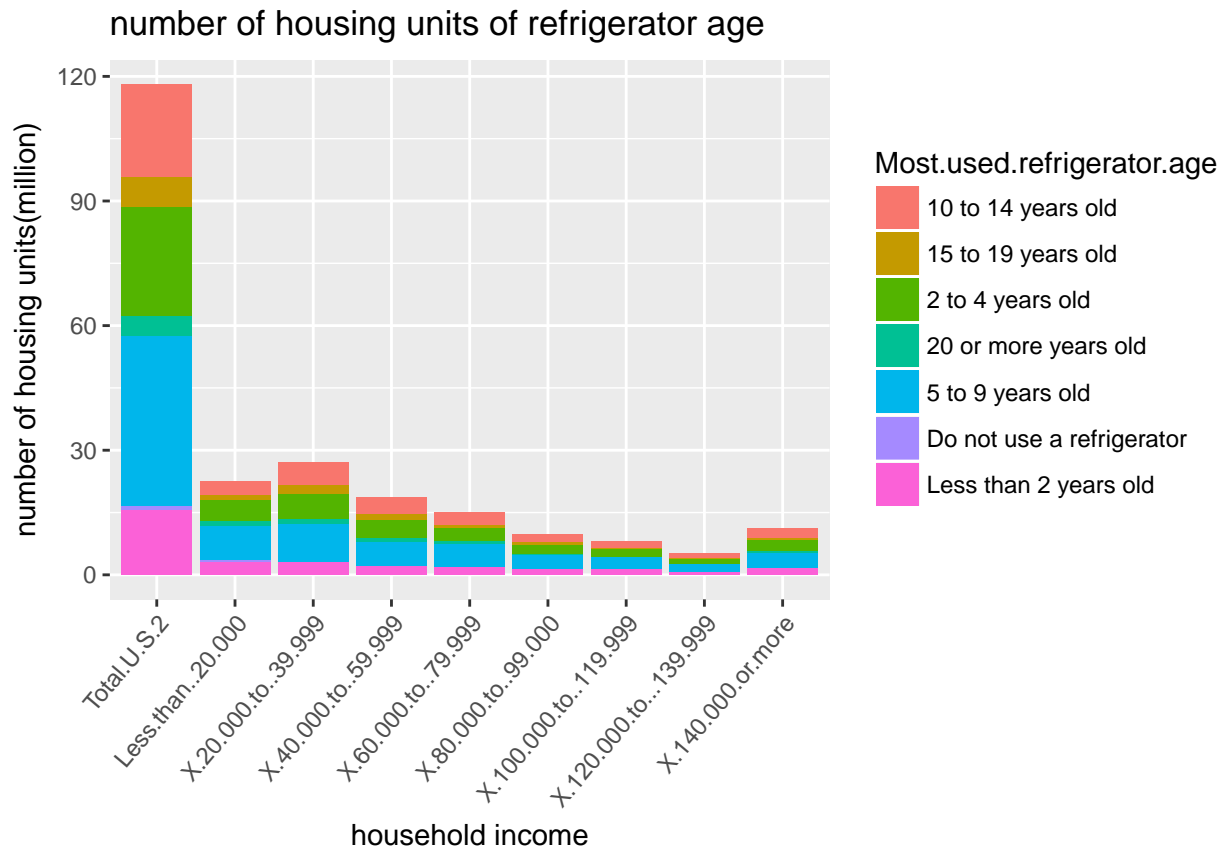
```
frisize[, 1] <- sapply(frisize[, 1], as.factor)
frisize[, 2] <- sapply(frisize[, 2], as.factor)
frisize[, 3] <- sapply(frisize[, 3], as.numeric)
ggfrisize <- ggplot() + geom_bar(aes(y = value, x = variable, fill = Most.used.refrigerator.size), data = fr isize)
ggfrisize
```

For the second histogram, In U.S., most of the households that have the different income use the medium sized fridge and then is the large sized fridge and nearly no household do not use a fridge.

EDA of fridge age

```
friage[ , 1] <- sapply(friage[ , 1], as.factor)
friage[ , 2] <- sapply(friage[ , 2], as.factor)
friage[ , 3] <- sapply(friage[ , 3], as.numeric)
ggfriage<-ggplot() + geom_bar(aes(y = value, x = variable, fill = Most.used.refrigerator.age), data = f
ggfriage
```



For the third histogram, In u.s., most of the households that have the different income have the fridges that are 5 to 9 years old. The second most is the fridges that are 2 to 4 years old. The households that have the different income have the least fridges that are less than 2 years old. The households that have the income of 20000 to 39999 have the highest household numbers that have fridges.

““

Conclusion

According to our analysis, most of the housing units recorded in the data owning at least one fridge, and the colder the region, the larger the household members all have positive impact. However, according to the diagrams, the most used fridge size is medium and used age is 5 to 9 years old. For the strategy we would suggest is that to have different marketing plans for different regions, a research on the households size and income distribution is necessary to be done before a new marketing launched.