MA415/615 EDA Project - Refrigerator

Xiaoqian Xue, Sibo Zhu, Li Liu, Yufei Lin, Danni Fu

2017/10/10

Introduction

In order to better design the strategy for marketing the refrigerators in U.S, we explore the EIA data in 2015 RECS Survey. In this project, we focus on three data sets: the Appliance by Household income (HC 3.5), the Refrigerator's size and type in the Northeast and Midwest regions (HC 3.7) and in the South and West regions (HC 3.8).

We assume that different regions may have different demands for the refrigerator size and type. Households living in the rich regions, for example New England or East Coast, may not need very large refrigerators since they are close to the market to buy fresh food frequently. And they may buy more expansive refrigerator with more function. On the contrary, households living in the middle or south may need large refrigerators with simple design.

In addition to analyzing the usage difference from the perspective of locations, we also analyze other data based on the assumption that people with different income level may have different demand for the refrigerator size and type. For example, people with higher income may prefer to use larger refrigerators because they can afford to buy various food and drinks to store in the refrigerators. On the other hand, people with lower income don't have that much money to spend on food or drinks so they only use medium or small refrigerators.

To see if there are certain differences, we use R to explore the data and plot to see the distributions of the EDA data and provide detailed analysis.

Exploring for different regions

We obtained the data from the Housing Appliances of February 27, 2017 in the Northeast and Midwest regions (HC3.7) and in the South and West regions (HC3.8).

```
# SUPPRESS GLOBAL WARNING
options(warn = -1)

# SET TO ONE DIGIT AFTER DECIMAL POINT
options(digits = 2)

# PREPARE PACKAGES
if (!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load("dplyr", "ggplot2", "tibble", "tidyr", "readxl", "reshape2")

#downloading data

download.file(
  "https://www.eia.gov/consumption/residential/data/2015/hc/hc3.7.xlsx",
  "NORTH_MID.xlsx", quiet = TRUE, mode = "wb")
```

```
download.file(
  "https://www.eia.gov/consumption/residential/data/2015/hc/hc3.8.xlsx",
  "SOUTH_WEST.xlsx", quiet = TRUE, mode = "wb")
```

To find the best marketing strategy, we want to know for northeast, Midwest, south and west regions, the number of housing units using refrigerator for different sizes, types and ages. Therefore we draw the data (i.e. most-used refrigerator size, most-used refrigerator type and most-used refrigerator age for New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain North, Mountain South and Pacific regions) from Table HC 3.7 and HC 3.8.

```
#selecting data
NM_size <- read_excel("NORTH_MID.xlsx", sheet = "data", range = "A108:H113",
                          col_names = FALSE, col_types = "text")

NM_type <- read_excel("NORTH_MID.xlsx", sheet = "data", range = "A115:H121",
                          col_names = FALSE, col_types = "text")

SW_size <- read_excel("SOUTH_WEST.xlsx", sheet = "data", range = "A111:K116",
                          col_names = FALSE, col_types = "text")

SW_type <- read_excel("SOUTH_WEST.xlsx", sheet = "data", range = "A118:K124",
                          col_names = FALSE, col_types = "text")
```

```
#create duplicate data for future use
NM_size_1 <- NM_size

NM_type_1 <- NM_type

SW_size_1 <- SW_size

SW_type_1 <- SW_type
```

```
#rename column's names
colnames(NM_size_1)      <- c("RFG_SIZE", "TTL_US", "TTL_NE",
                          "N_ENG", "MID_ATL", "TTL_MID_WEST",
                          "EN_CENT", "WN_CENT")

colnames(NM_type_1)      <- c("RFG_TYPE", "TTL_US", "TTL_NE",
                          "N_ENG", "MID_ATL", "TTL_MID_WEST",
                          "EN_CENT", "WN_CENT")

colnames(SW_size_1)      <- c("RFG_SIZE", "TTL_US", "TTL_SOUTH",
                          "S_ATL", "ES_CENT", "WS_CENT",
                          "TTL_WEST", "TTL_MOUNT", "MOUNT_N",
                          "MOUNT_SOUTH", "PACIF")

colnames(SW_type_1)      <- c("RFG_TYPE", "TTL_US", "TTL_SOUTH",
                          "S_ATL", "ES_CENT", "WS_CENT",
                          "TTL_WEST", "TTL_MOUNT", "MOUNT_N",
                          "MOUNT_SOUTH", "PACIF")
```

```
#Saving total column for future use
NM_size_TTL <- within(NM_size_1,rm(TTL_US,N_ENG,MID_ATL,EN_CENT,WN_CENT))
NM_type_TTL <- within(NM_type_1,rm(TTL_US,N_ENG,MID_ATL,EN_CENT,WN_CENT))
```

```
SW_size_TTL <- within(SW_size_1,rm(TTL_US,S_ATL,ES_CENT,WS_CENT,TTL_MOUNT,MOUNT_N,MOUNT_SOUTH,PACIF))
SW_type_TTL <- within(SW_type_1,rm(TTL_US,S_ATL,ES_CENT,WS_CENT,TTL_MOUNT,MOUNT_N,MOUNT_SOUTH,PACIF))
```

```
#Drop the Unnecessary Column

NM_size_2 <- within(NM_size_1, rm(TTL_US,TTL_NE,TTL_MID_WEST))
NM_type_2 <- within(NM_type_1, rm(TTL_US,TTL_NE,TTL_MID_WEST))
SW_size_2 <- within(SW_size_1, rm(TTL_US,TTL_SOUTH,TTL_WEST,TTL_MOUNT))
SW_type_2 <- within(SW_type_1, rm(TTL_US,TTL_SOUTH,TTL_WEST,TTL_MOUNT))
```

```
# COERCE FRIDGE TYPE COLUMN TO FACTOR
NM_size_2$RFG_SIZE    <- as.factor(NM_size_2$RFG_SIZE)
NM_type_2$RFG_TYPE    <- as.factor(NM_type_2$RFG_TYPE)
SW_size_2$RFG_SIZE    <- as.factor(SW_size_2$RFG_SIZE)
SW_type_2$RFG_TYPE    <- as.factor(SW_type_2$RFG_TYPE)

NM_size_TTL$RFG_SIZE    <- as.factor(NM_size_TTL$RFG_SIZE)
NM_type_TTL$RFG_TYPE    <- as.factor(NM_type_TTL$RFG_TYPE)
SW_size_TTL$RFG_SIZE    <- as.factor(SW_size_TTL$RFG_SIZE)
SW_type_TTL$RFG_TYPE    <- as.factor(SW_type_TTL$RFG_TYPE)
```

```
#cleaning data, mark unknown data as N/A
NM_size_2[, 2:5]    <- sapply(NM_size_2[, 2:5], as.numeric)
NM_type_2[, 2:5]    <- sapply(NM_type_2[, 2:5], as.numeric)
SW_size_2[, 2:7]    <- sapply(SW_size_2[, 2:7], as.numeric)
SW_type_2[, 2:7]    <- sapply(SW_type_2[, 2:7], as.numeric)
NM_size_TTL[, 2:3]    <- sapply(NM_size_TTL[, 2:3], as.numeric)
NM_type_TTL[, 2:3]    <- sapply(NM_type_TTL[, 2:3], as.numeric)

SW_size_TTL[, 2:3]    <- sapply(SW_size_TTL[, 2:3], as.numeric)
SW_type_TTL[, 2:3]    <- sapply(SW_type_TTL[, 2:3], as.numeric)
```

```
#rename columns
colnames(NM_size_2)      <- c("RFG_SIZE",
                              "N_ENG", "MID_ATL",
                              "EN_CENT", "WN_CENT")

colnames(NM_type_2)      <- c("RFG_TYPE",
                              "N_ENG", "MID_ATL",
                              "EN_CENT", "WN_CENT")

colnames(SW_size_2)      <- c("RFG_SIZE",
                              "S_ATL", "ES_CENT", "WS_CENT",
                              "MOUNT_N",
                              "MOUNT_SOUTH", "PACIF")

colnames(SW_type_2)      <- c("RFG_TYPE",
                              "S_ATL", "ES_CENT", "WS_CENT",
                              "MOUNT_N",
                              "MOUNT_SOUTH", "PACIF")
```

```
#Melting Data
NM_size_3 <- as_tibble(melt(NM_size_2, id = 1))
NM_type_3 <- as_tibble(melt(NM_type_2, id = 1))
SW_size_3 <- as_tibble(melt(SW_size_2, id = 1))
```

```
SW_type_3 <- as_tibble(melt(SW_type_2, id = 1))
NM_size_TTL <- as_tibble(melt(NM_size_TTL, id = 1))
NM_type_TTL <- as_tibble(melt(NM_type_TTL, id = 1))
SW_size_TTL <- as_tibble(melt(SW_size_TTL, id = 1))
SW_type_TTL <- as_tibble(melt(SW_type_TTL, id = 1))
```

```
#Update column's name
names(NM_size_3)[2]    <- "Region"
names(NM_type_3)[2]    <- "Region"
names(SW_size_3)[2]    <- "Region"
names(SW_type_3)[2]    <- "Region"
names(NM_size_TTL)[2]    <- "Region"
names(NM_type_TTL)[2]    <- "Region"
names(SW_size_TTL)[2]    <- "Region"
names(SW_type_TTL)[2]    <- "Region"

####

names(NM_size_3)[3]    <- "Number of Unit"
names(NM_type_3)[3]    <- "Number of Unit"
names(SW_size_3)[3]    <- "Number of Unit"
names(SW_type_3)[3]    <- "Number of Unit"
names(NM_size_TTL)[3]    <- "Number of Unit"
names(NM_type_TTL)[3]    <- "Number of Unit"
names(SW_size_TTL)[3]    <- "Number of Unit"
names(SW_type_TTL)[3]    <- "Number of Unit"
```

```
#getting sum of data
NM_size_SUM <- aggregate(NM_size_3$`Number of Unit`,
                         by = list(category = NM_size_3$`Region`),
                         FUN=sum, na.rm = TRUE)

NM_type_SUM <- aggregate(NM_type_3$`Number of Unit`,
                         by = list(category = NM_type_3$`Region`),
                         FUN=sum, na.rm = TRUE)

SW_size_SUM <- aggregate(SW_size_3$`Number of Unit`,
                         by = list(category = SW_size_3$`Region`),
                         FUN=sum, na.rm = TRUE)

SW_type_SUM <- aggregate(SW_type_3$`Number of Unit`,
                         by = list(category = SW_type_3$`Region`),
                         FUN=sum, na.rm = TRUE)
```

```
#Rename Columns
names(NM_size_SUM)[1] <- "Region"
names(NM_size_SUM)[2] <- "Total Number of Unit"

names(NM_type_SUM)[1] <- "Region"
names(NM_type_SUM)[2] <- "Total Number of Unit"

names(SW_size_SUM)[1] <- "Region"
names(SW_size_SUM)[2] <- "Total Number of Unit"
```

```r
names(SW_type_SUM)[1] <- "Region"
names(SW_type_SUM)[2] <- "Total Number of Unit"
```

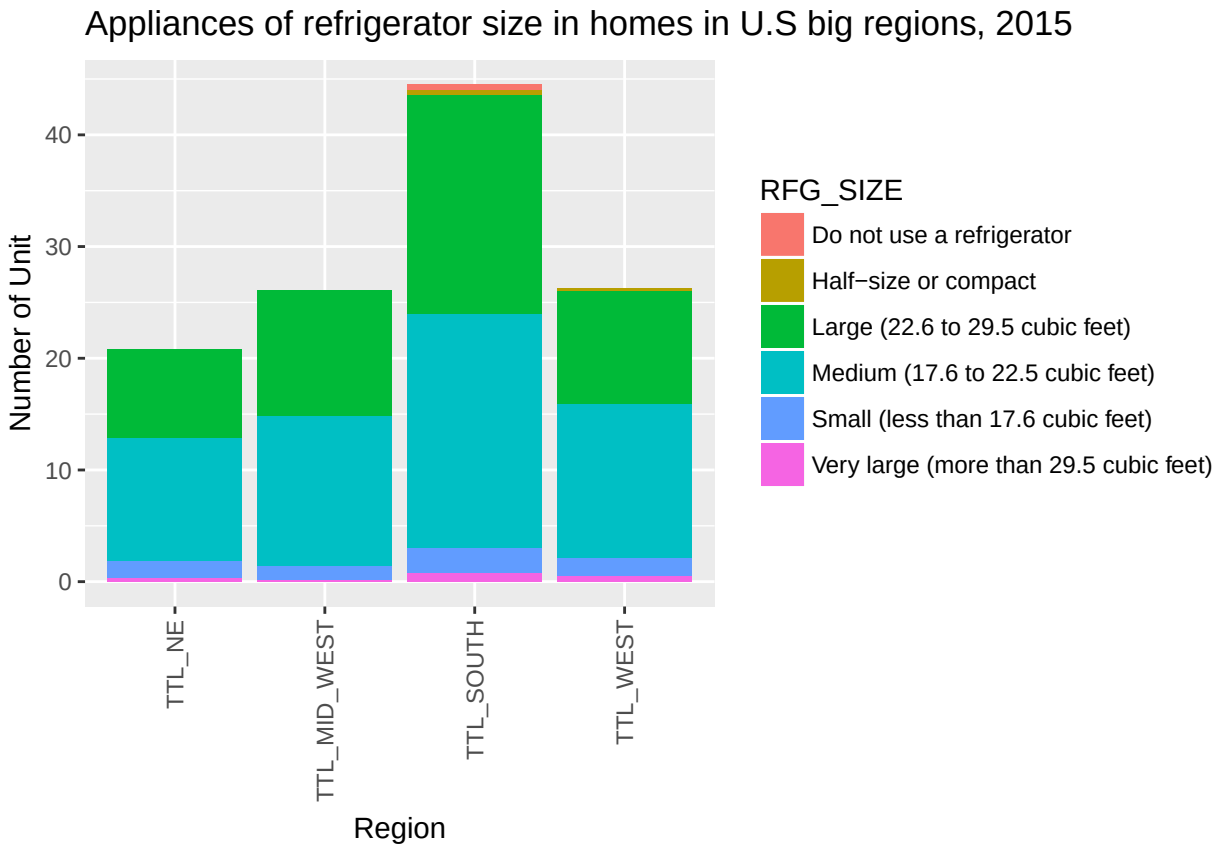
```r
#Merge Sum with Original Data
TTL_NM_size <- merge(x = NM_size_3, y = NM_size_SUM, by = "Region", all = TRUE)
names(TTL_NM_size)[3] <- "Number of Unit"
names(TTL_NM_size)[4] <- "Total Number of Unit by Region"
TTL_NM_size$Percentage <- 100*(TTL_NM_size$`Number of Unit` /TTL_NM_size$`Total Number of Unit by Regio
rm(list = ls(pattern = "^T_"))

TTL_NM_type <- merge(x = NM_type_3, y = NM_type_SUM, by = "Region", all = TRUE)
names(TTL_NM_type)[3] <- "Number of Unit"
names(TTL_NM_type)[4] <- "Total Number of Unit by Region"
TTL_NM_type$Percentage <- 100*(TTL_NM_type$`Number of Unit` /TTL_NM_type$`Total Number of Unit by Regio
rm(list = ls(pattern = "^T_"))

TTL_SW_size <- merge(x = SW_size_3, y = SW_size_SUM, by = "Region", all = TRUE)
names(TTL_SW_size)[3] <- "Number of Unit"
names(TTL_SW_size)[4] <- "Total Number of Unit by Region"
TTL_SW_size$Percentage <- 100*(TTL_SW_size$`Number of Unit` /TTL_SW_size$`Total Number of Unit by Regio
rm(list = ls(pattern = "^T_"))

TTL_SW_type <- merge(x = SW_type_3, y = SW_type_SUM, by = "Region", all = TRUE)
names(TTL_SW_type)[3] <- "Number of Unit"
names(TTL_SW_type)[4] <- "Total Number of Unit by Region"
TTL_SW_type$Percentage <- 100*(TTL_SW_type$`Number of Unit` /TTL_SW_type$`Total Number of Unit by Regio
rm(list = ls(pattern = "^T_"))

TTL_size <- merge(x = NM_size_TTL, y = SW_size_TTL, by = c("RFG_SIZE","Region","Number of Unit"), all =
TTL_type <- merge(x = NM_type_TTL, y = SW_type_TTL, by = c("RFG_TYPE","Region","Number of Unit"), all =
rm(list = ls(pattern = "^T_"))
```

```r
#Start plotting
PLOT_NM_SIZE_PERC <- ggplot(data = TTL_NM_size, aes(x = `Region`, y = Percentage)) +
                geom_col(aes(fill = `RFG_SIZE`)) +
                labs(title = "Percentage of Northeast and Midwest regions, 2015") +
                coord_flip()

PLOT_NM_SIZE <- ggplot(data = TTL_NM_size, aes(x = `Region`, y = `Number of Unit`)) +
             geom_col(aes(fill = `RFG_SIZE`)) +
             theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
             labs(title = "Appliances in homes in the Northeast and Midwest regions, 2015")

PLOT_NM_TYPE_PERC <- ggplot(data = TTL_NM_type, aes(x = `Region`, y = Percentage)) +
                geom_col(aes(fill = `RFG_TYPE`)) +
                labs(title = "Percentage of Northeast and Midwest regions, 2015") +
                coord_flip()

PLOT_NM_TYPE <- ggplot(data = TTL_NM_type, aes(x = `Region`, y = `Number of Unit`)) +
             geom_col(aes(fill = `RFG_TYPE`)) +
             theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
             labs(title = "Appliances in homes in the Northeast and Midwest regions, 2015")

############
```

```
PLOT_SW_SIZE_PERC <- ggplot(data = TTL_SW_size, aes(x = `Region`, y = Percentage)) +
                geom_col(aes(fill = `RFG_SIZE`)) +
                labs(title = "Percentage of South and West regions, 2015") +
                coord_flip()

PLOT_SW_SIZE <- ggplot(data = TTL_SW_size, aes(x = `Region`, y = `Number of Unit`)) +
             geom_col(aes(fill = `RFG_SIZE`)) +
             theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
             labs(title = "Appliances in homes in the South and West regions, 2015")

PLOT_SW_TYPE_PERC <- ggplot(data = TTL_SW_type, aes(x = `Region`, y = Percentage)) +
                geom_col(aes(fill = `RFG_TYPE`)) +
                labs(title = "Percentage of South and West regions, 2015") +
                coord_flip()

PLOT_SW_TYPE <- ggplot(data = TTL_SW_type, aes(x = `Region`, y = `Number of Unit`)) +
             geom_col(aes(fill = `RFG_TYPE`)) +
             theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
             labs(title = "Appliances in homes in the South and West regions, 2015")

######

PLOT_SIZE_TTL <- ggplot(data = TTL_size, aes(x = `Region`, y = `Number of Unit`)) +
             geom_col(aes(fill = `RFG_SIZE`)) +
             theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
             labs(title = "Appliances of refrigerator size in homes in U.S big regions, 2015")

PLOT_TYPE_TTL <- ggplot(data = TTL_type, aes(x = `Region`, y = `Number of Unit`)) +
             geom_col(aes(fill = `RFG_TYPE`)) +
             theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
             labs(title = "Appliances of refrigerator type in homes in U.S big regions, 2015")
```

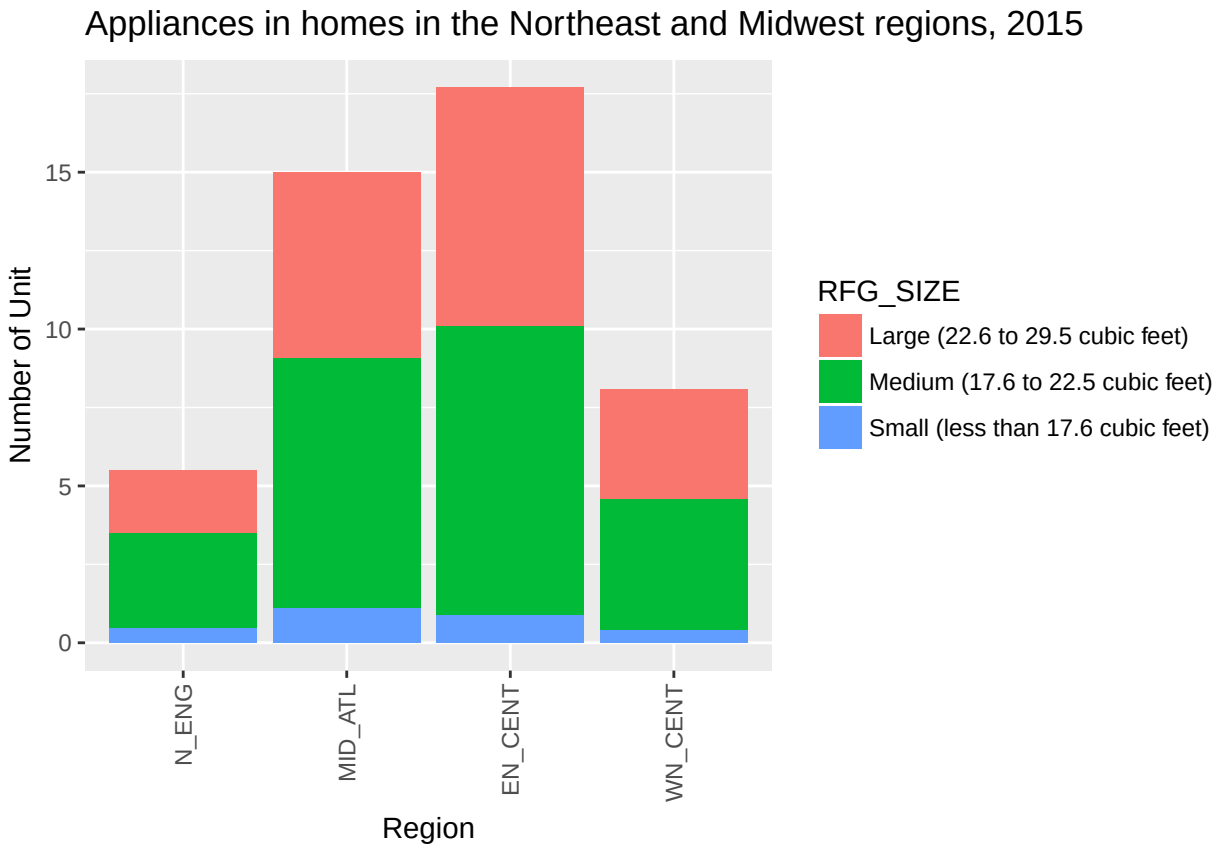
First, we compared the appliances of refrigerator size in homes in overall U.S regions by using the ggplot. The refrigerator sizes include small (less than 17.6 cubic feet), medium (17.6 to 22.5 cubic feet), large (22.6 to 29.5 cubic feet) and very large (more than 29.5 cubic feet). The vertical axis is the number of housing unit and the horizontal axis is four regions, total northeast, total Midwest plus Middle Atlantic, Total South and Total West. It is clearly shown in the ggplot that most household in the U.S. use the medium and large size refrigerator. There is no big difference between different regions in the size of refrigerator used by households. Only a small amount of household uses the small size and approximately none of the household use very large or do not use a refrigerator.

```
PLOT_SIZE_TTL
```



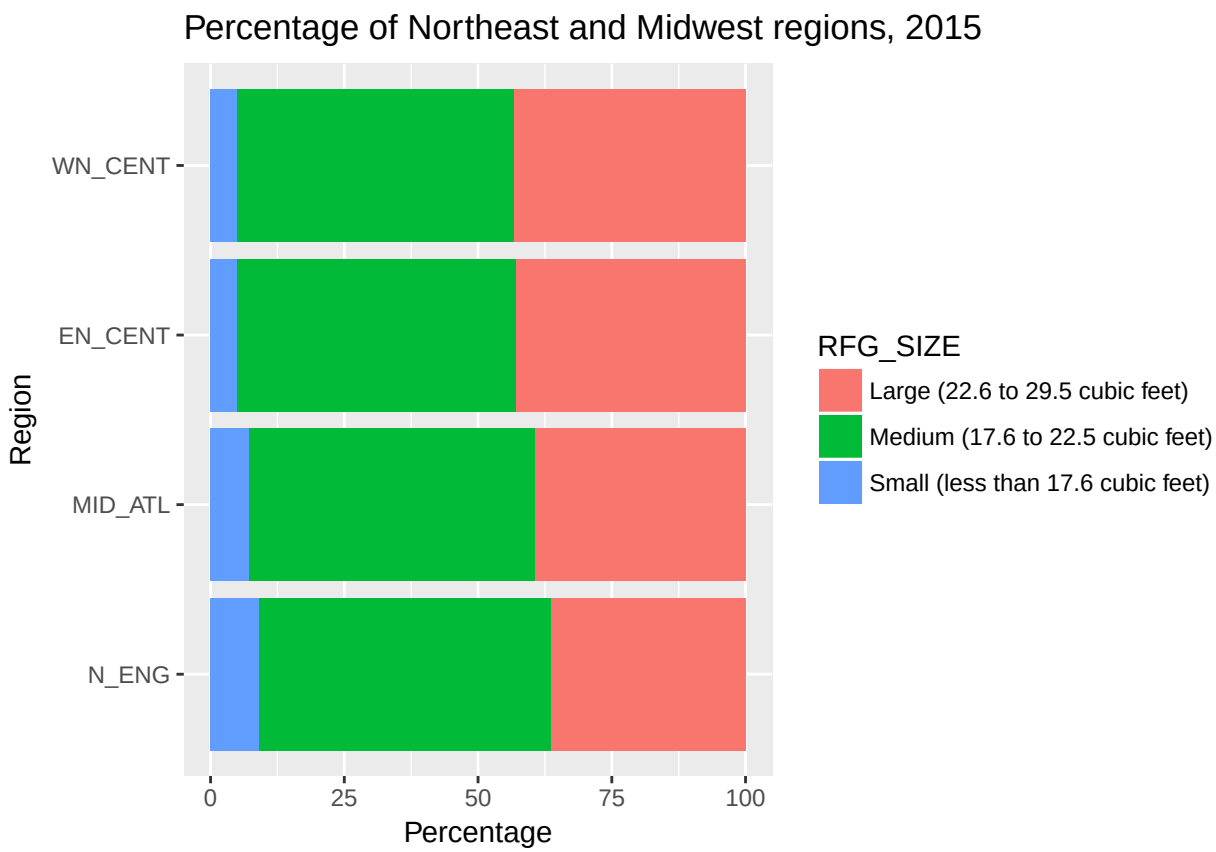
Next, we focus on the total northeast, total Midwest regions, specifically New England, Middle Atlantic, East North Central and West North Central. We find out that all surveyed households us a refrigerator and none of the surveyed households use half-size or compact refrigerator or very large size refrigerator. Medium and large size refrigerator are more popular in those 4 regions compared to the small size.

```
PLOT_NM_SIZE
```



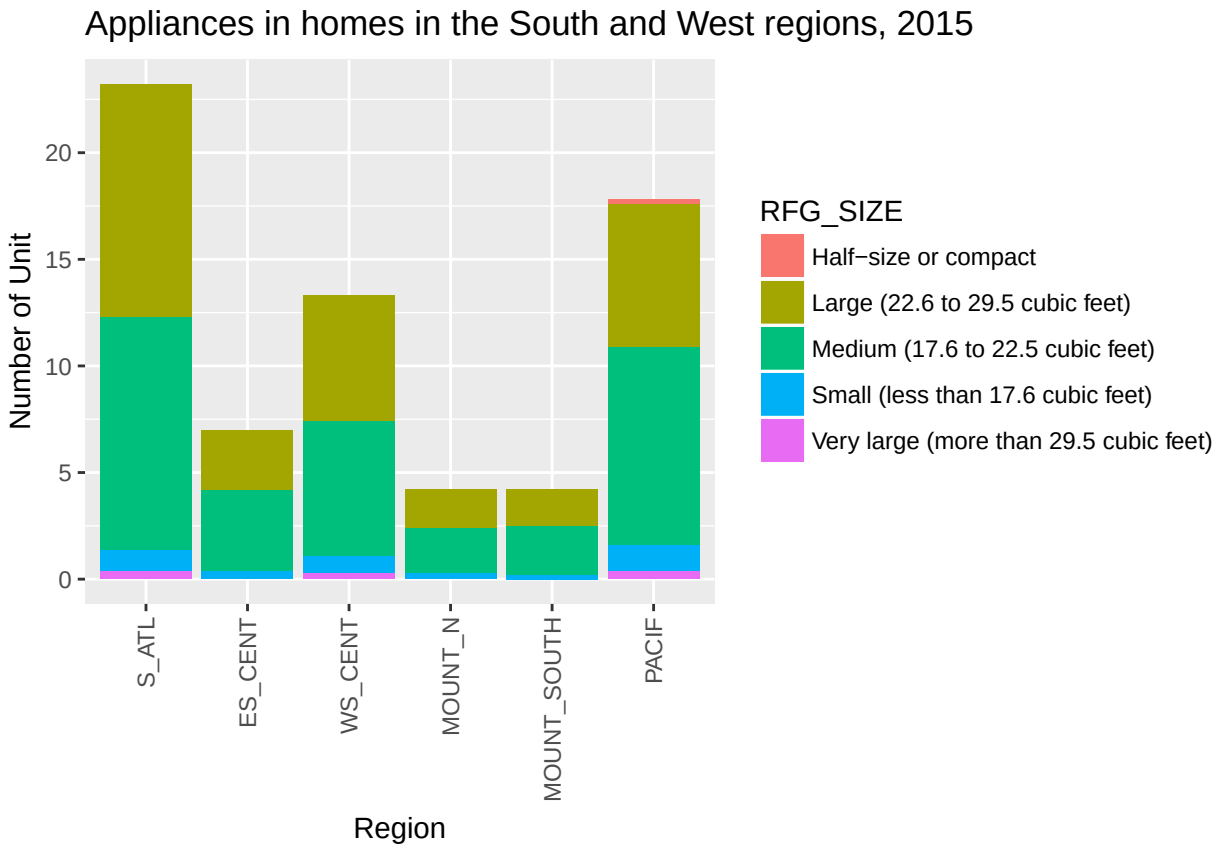
To find out which size of refrigerator is the most popular one in the each part of the Northeast and Midwest regions, we calculate the percentage of the number of households using specific size by the total number of households in different regions. Then we use the ggplot to compare the difference. We find out there is no difference between West North Central and East North Central that the medium size refrigerator is more popular than the large size one. Large size refrigerator tends to be more dominate in the Middle Atlantic and New England. Therefore, we should market the medium size refrigerator in the West North Central and East North Central, and the large size refrigerator in the Middle Atlantic and New England.

```
PLOT_NM_SIZE_PERC
```



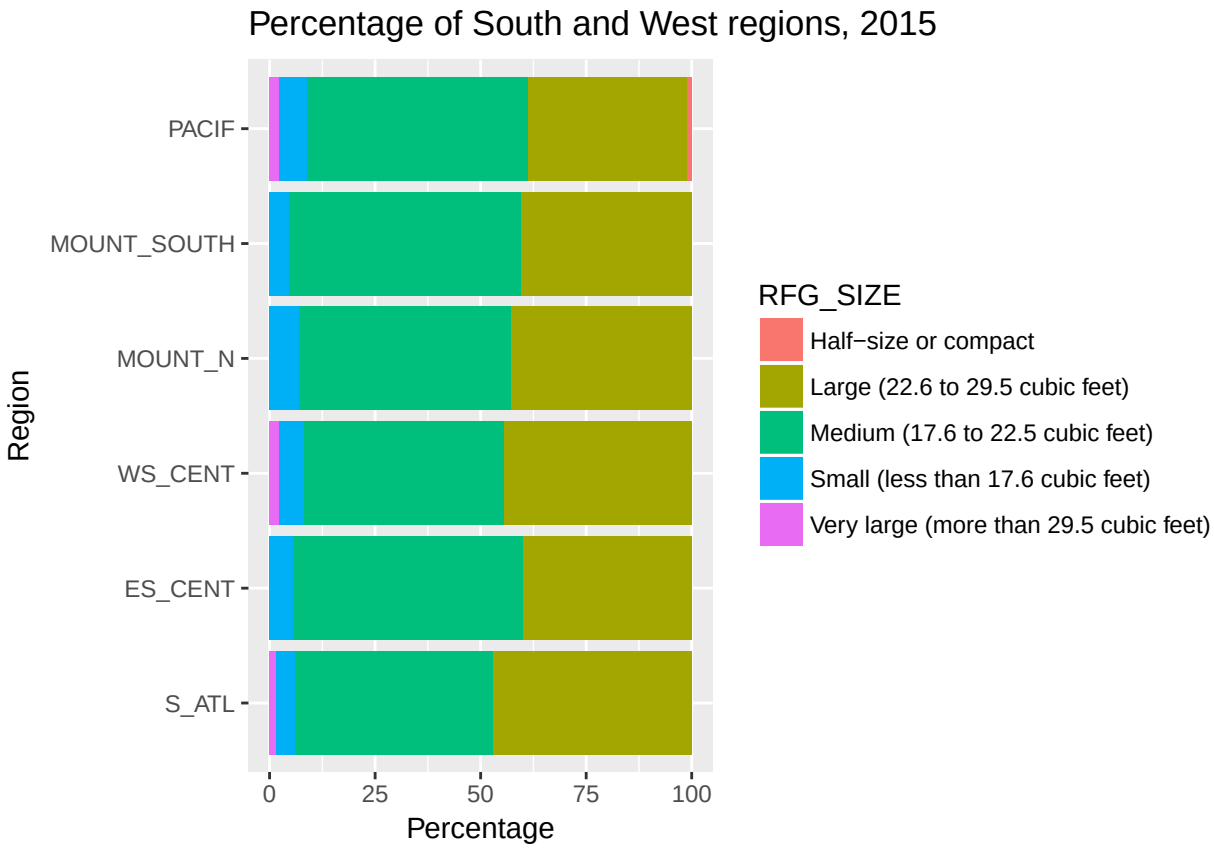
Then we focus on the South and West regions, specifically South Atlantic, East South Central, West South Central, Mountain North, Mountain South and Pacific. Only tiny number of households use the half-size or compact, small size or very large size refrigerator. In the South Atlantic and West Central, the large and medium size are both popular. However, in the East South Central and Pacific, medium size refrigerator seems more popular. The ggplot also shows that the distribution of Mountain North and South are similar in which the large size and the medium size refrigerator equally dominate.

```
PLOT_SW_SIZE
```



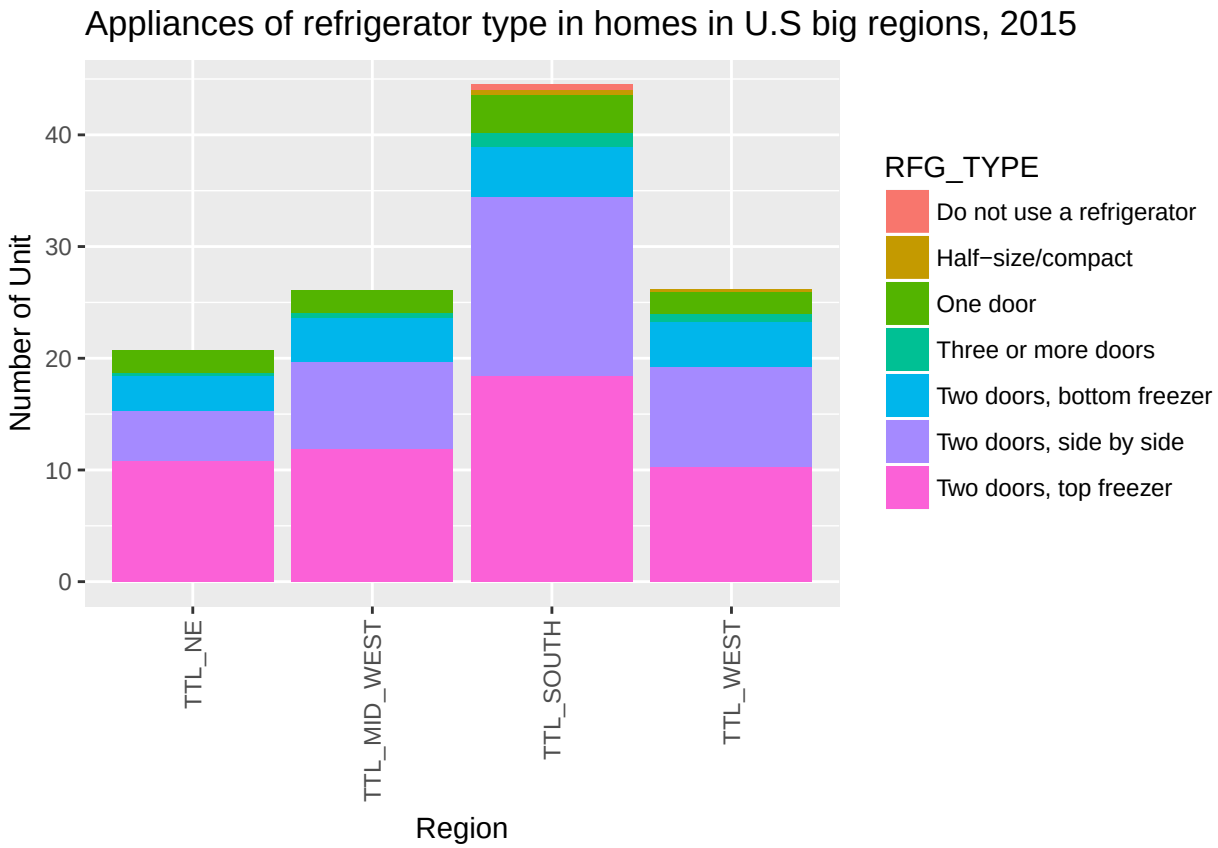
To get a more accurate information to confirm the previous observation, we also calculate the percentage of the number of households using specific size by the total number of households in different regions. Then we use the ggplot to compare the difference in the South and West Regions. We find out medium size refrigerator is the most popular in the Pacific, Mountain South, Mountain North and East Central. For West Central and South Atlantic, the medium size and large size refrigerator are equally popular. Therefore, we can market the medium size refrigerator in all South and West regions and the large size refrigerator in West Central and South Atlantic.

```
PLOT_SW_SIZE_PERC
```



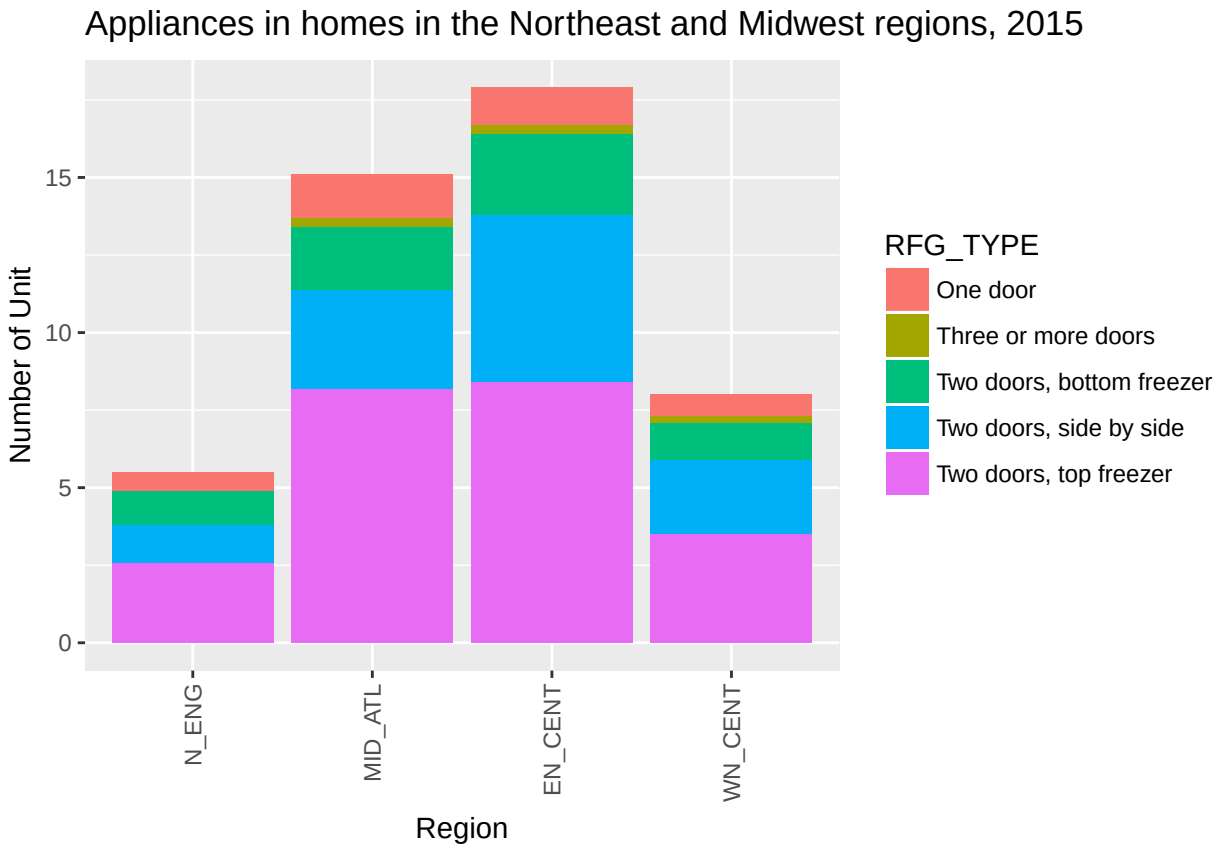
To better market, we compared the appliances of refrigerator types in homes in overall U.S regions by using the ggplot. The refrigerator types vary in half-size/compact, one door, three or more doors, two doors with bottom freezer, two doors with side by side, two doors with top freezer. The vertical axis is the number of housing unit and the horizontal axis is four regions, total northeast, total Midwest plus Middle Atlantic, Total South and Total West. It is clearly shown in the ggplot that most household in the U.S. use two doors refrigerators with top freezer and two doors refrigerator with side by side. Only few of the household does not use a refrigerator. There is no big difference between different regions in the types of refrigerator used by households.

```
PLOT_TYPE_TTL
```



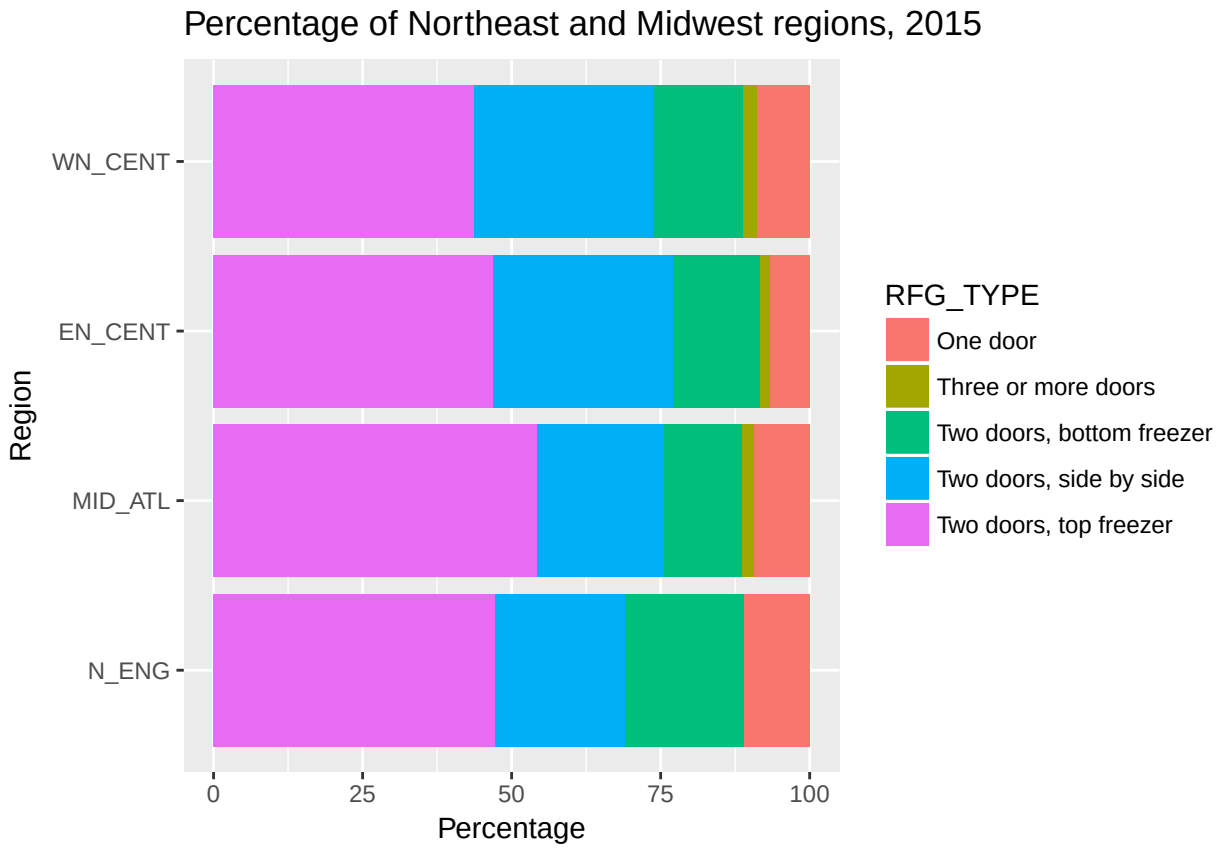
First, we study on the appliances of refrigerator types in the total Northeast and total Midwest regions, consisting of New England, Middle Atlantic, East North Central and West North Central. It is shown that all surveyed households have a refrigerator, only few households in all region use a one door refrigerator, and none of them use a half-size or compact refrigerator. Two doors refrigerators with top freezer and two doors refrigerators with side by side occupy the majority of households' appliance in all five regions in Northeast and Midwest regions compare to other types.

```
PLOT_NM_TYPE
```



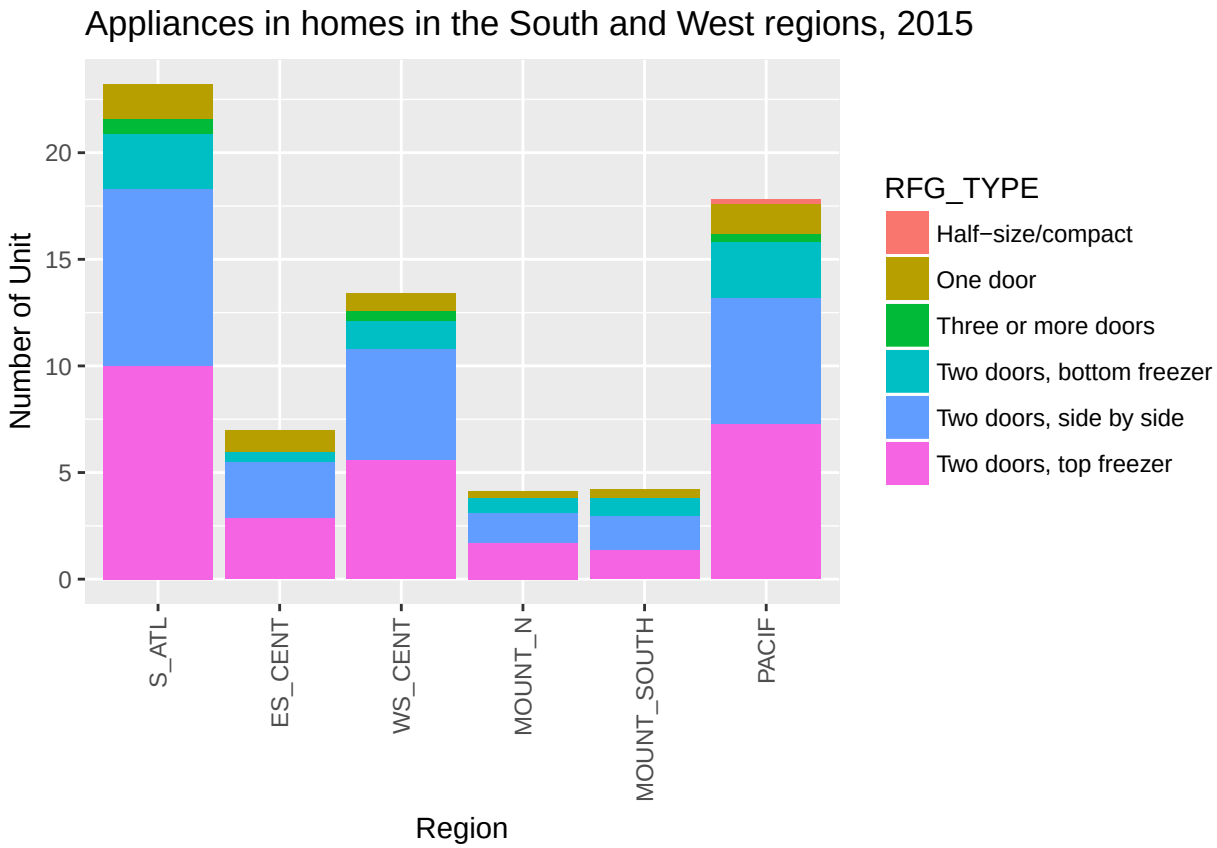
To make a further conclusion, the percentage of the refrigerator types of each region is calculated by dividing the number of households using each type from the total number of appliance in specific region. Then we use ggplot to get a clearer version of the distribution of each type in specific regions. It is shown that majority of those households in Northeast and Midwest regions uses two doors refrigerators with top freezers. Tiny amount of households in Middle Atlantic, East North Central and West North Central uses the three doors refrigerators, and none of those from New England use three doors'. And there is no big difference in the distribution of different types among all the regions.

```
PLOT_NM_TYPE_PERC
```



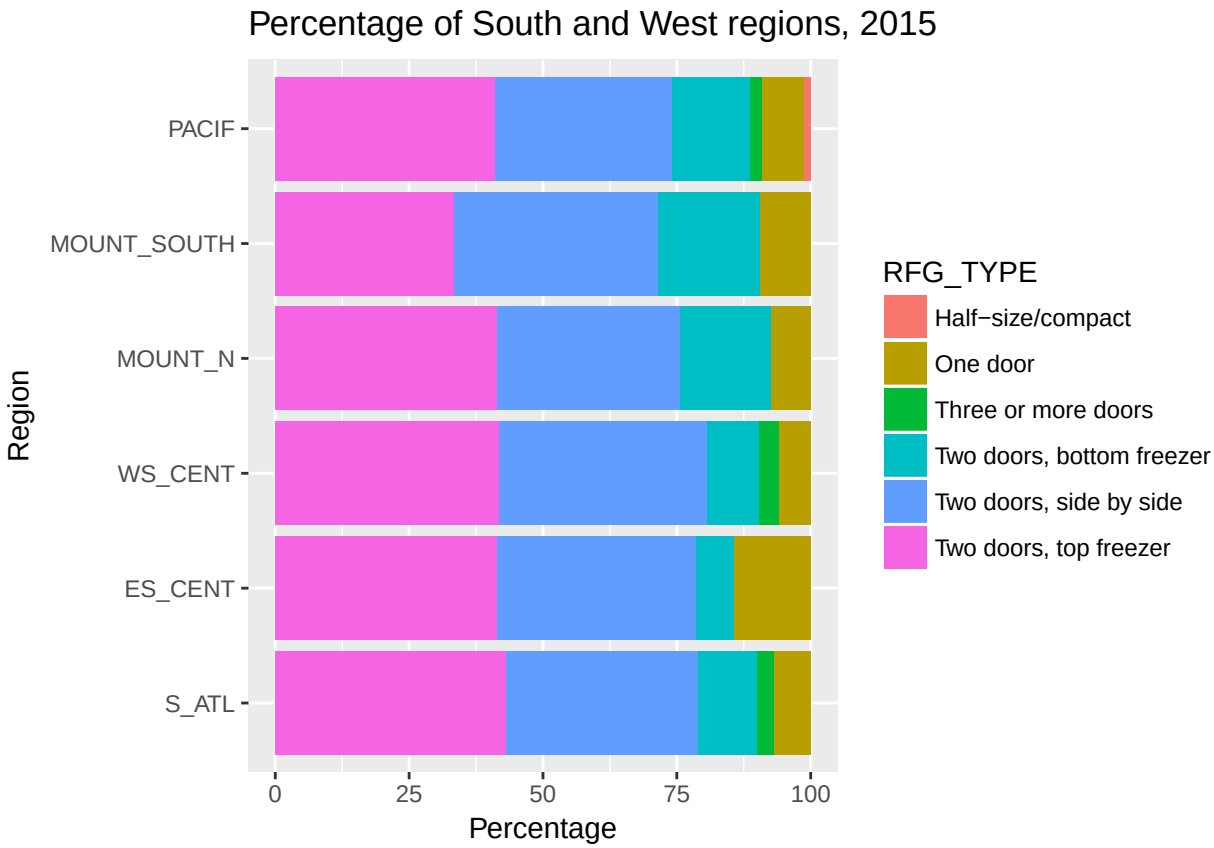
Next, we focus on the South and West regions, consisting of 6 sub regions, South Atlantic, East South Central, West South Central, Mountain North, Mountain South and Pacific regions. According to the ggplot, we can see that the most household in South and West prefer the two door refrigerator with top freezer and two doors refrigerators with side by side. Only tiny amount of household in Pacific region does not use a refrigerator. And little amount of the half-size or compact refrigerators is used by all those regions in South and West, and few of the household in all those South and West regions use the one door refrigerators.

```
PLOT_SW_TYPE
```



To confirm our analysis, the percentage of each refrigerator types used by households in different regions are calculated by the total number of household in different regions of the South and West. Ggplot is effectively used to compare the different preference of refrigerator type of all six regions. We found out that the majority of the household of each region uses the two doors refrigerators with top freezers and two doors refrigerator with side by side. Thus, we can conclude that two doors refrigerators with top freezers and two doors refrigerators with side by side should be marketed in all South and West regions.

```
PLOT_SW_TYPE_PERC
```



Conclusion

In general, medium size and two doors refrigerator with top freezers is the most popular size throughout U.S, especially in the West North Central and East North Central, in all South and West regions. Also, the two doors refrigerators with top freezers can also be marketed in all South and West regions. The large size refrigerator can also be the another target for marketing, since it is more popular than the medium size in the Middle Atlantic , New England , West Central and South Atlantic.

Exploring for different Income Levels

```
#Income Analysis Part
#downloading data

download.file(
  "https://www.eia.gov/consumption/residential/data/2015/hc/hc3.5.xlsx",
  "INCOME.xlsx", quiet = TRUE, mode = "wb")

#selecting data
sizeData <- read_excel("INCOME.xlsx", sheet = "data", range = "A108:J113",
                       col_names = FALSE, col_types = "text")

typeData <- read_excel("INCOME.xlsx", sheet = "data", range = "A115:J121",
                       col_names = FALSE, col_types = "text")
```

```r
#create duplicate data for future use
sizeData_1 <- sizeData

typeData_1 <- typeData


colnames(sizeData_1)      <- c("RFG_SIZE", "TTL_US", "LESS_20T",
                           "20_39T", "40_59T", "60_79T",
                           "80_99T", "100_119T","120_139T","140T_MORE")

colnames(typeData_1)      <- c("RFG_TYPE", "TTL_US", "LESS_20T",
                           "20_39T", "40_59T", "60_79T",
                           "80_99T", "100_119T","120_139T","140T_MORE")
```

```r
#Drop the Unnecessary Column

sizeData_2 <- within(sizeData_1, rm(TTL_US))
typeData_2 <- within(typeData_1, rm(TTL_US))



# COERCE FRIDGE TYPE COLUMN TO FACTOR
sizeData_2$RFG_SIZE    <- as.factor(sizeData_2$RFG_SIZE)
typeData_2$RFG_TYPE    <- as.factor(typeData_2$RFG_TYPE)



#cleaning data, mark unknown data as N/A
sizeData_2[, 2:9]    <- sapply(sizeData_2[, 2:9], as.numeric)
typeData_2[, 2:9]    <- sapply(typeData_2[, 2:9], as.numeric)
```

```r
#Melting Data
sizeData_3 <- as_tibble(melt(sizeData_2, id = 1))
typeData_3 <- as_tibble(melt(typeData_2, id = 1))

#Update column's name
names(sizeData_3)[2]    <- "Income Range"
names(sizeData_3)[3]    <- "Number of Unit"
names(typeData_3)[2]    <- "Income Range"
names(typeData_3)[3]    <- "Number of Unit"
```

We first select the data related to the refrigerators features that we are interested in from the income dataset–size and type and organize the data.

```r
#getting sum of data
sizeData_SUM <- aggregate(sizeData_3$`Number of Unit`,
                          by = list(category = sizeData_3$`Income Range`),
                          FUN=sum, na.rm = TRUE)

typeData_SUM <- aggregate(typeData_3$`Number of Unit`,
                          by = list(category = typeData_3$`Income Range`),
                          FUN=sum, na.rm = TRUE)
```

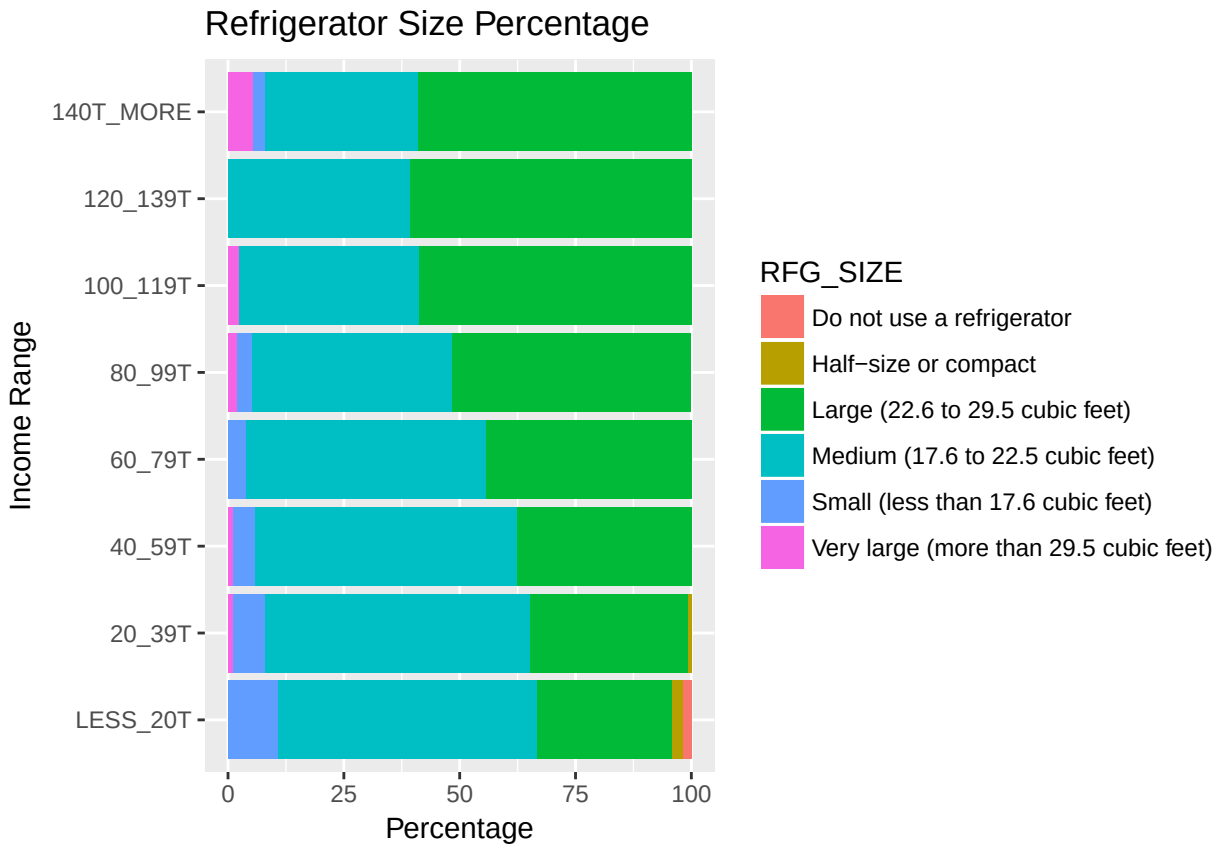
```r
#Rename Columns
names(sizeData_SUM)[1] <- "Income Range"
names(sizeData_SUM)[2] <- "Total Number of Unit"

names(typeData_SUM)[1] <- "Income Range"
names(typeData_SUM)[2] <- "Total Number of Unit"



#Merge Sum with Original Data
sizeDataWithTotal <- merge(x = sizeData_3, y = sizeData_SUM, by = "Income Range", all = TRUE)
names(sizeDataWithTotal)[4] <- "Total Number of Unit by Income Range"
sizeDataWithTotal$Percentage <- 100*(sizeDataWithTotal$`Number of Unit` /sizeDataWithTotal$`Total Numbe
rm(list = ls(pattern = "^T_"))

typeDataWithTotal <- merge(x = typeData_3, y = typeData_SUM, by = "Income Range", all = TRUE)
names(typeDataWithTotal)[4] <- "Total Number of Unit by Income Range"
typeDataWithTotal$Percentage <- 100*(typeDataWithTotal$`Number of Unit` /typeDataWithTotal$`Total Numbe
rm(list = ls(pattern = "^T_"))
```

Here we prepare the data to do the plotting for groups of people with different income ranges.

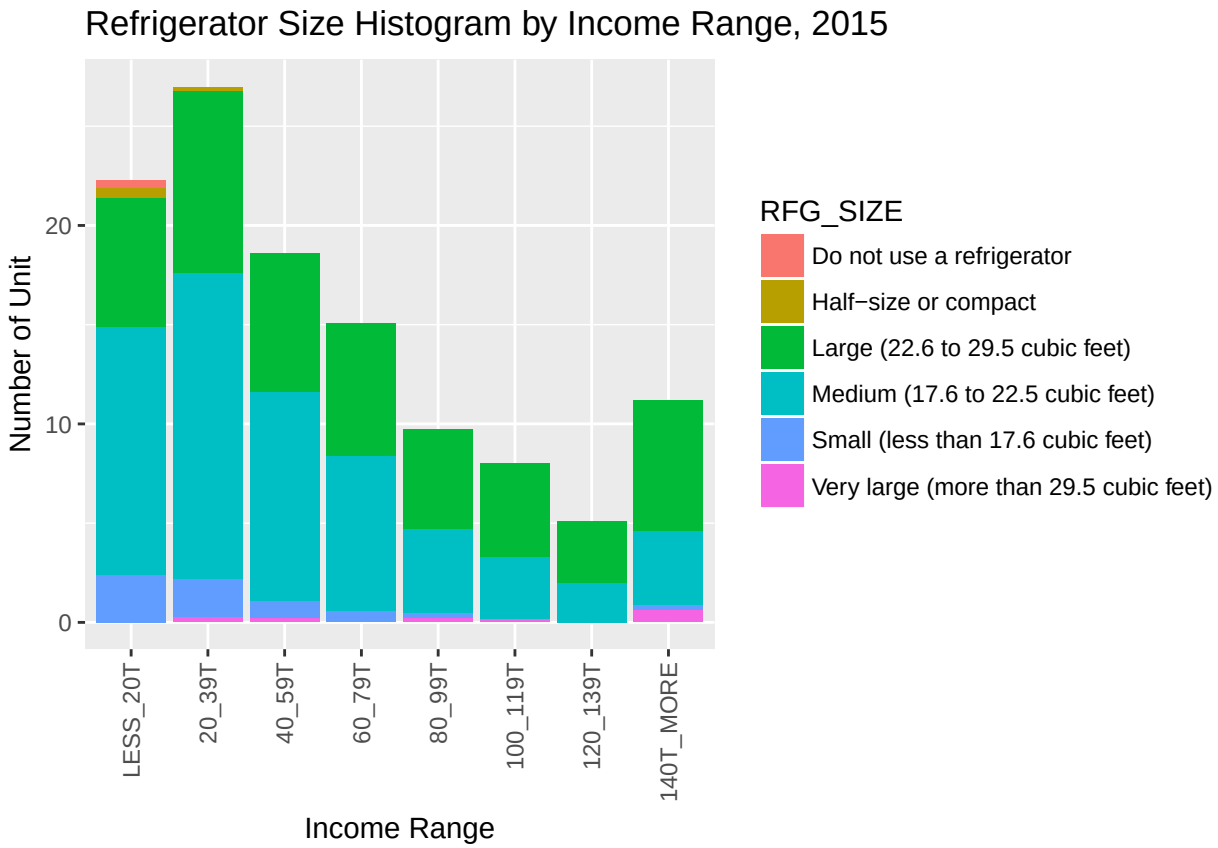
```r
#Start plotting
PLOT_SIZE_PERC <- ggplot(data = sizeDataWithTotal, aes(x = `Income Range`, y = Percentage)) +
                  geom_col(aes(fill = `RFG_SIZE`)) +
                  labs(title = "Refrigerator Size Percentage") +
                  coord_flip()

PLOT_SIZE <- ggplot(data = sizeDataWithTotal, aes(x = `Income Range`, y = `Number of Unit`)) +
             geom_col(aes(fill = `RFG_SIZE`)) +
             theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
             labs(title = "Refrigerator Size Histogram by Income Range, 2015")

PLOT_TYPE_PERC <- ggplot(data = typeDataWithTotal, aes(x = `Income Range`, y = Percentage)) +
                  geom_col(aes(fill = `RFG_TYPE`)) +
                  labs(title = "Refrigerator Type Percentage") +
                  coord_flip()

PLOT_TYPE <- ggplot(data = typeDataWithTotal, aes(x = `Income Range`, y = `Number of Unit`)) +
             geom_col(aes(fill = `RFG_TYPE`)) +
             theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
             labs(title = "Refrigerator Type Histogram by Income Range, 2015")
```

A common way to connect marketing strategy with income range is to market according to the living standard of different areas. For example, if we know that most housing units with high income use large refrigerators, we will mark large refrigerators as our main products in the areas with high living standard. The reason is that housing units with higher income usually have higher requirement of living standard. There are different ways to define living standard of an area. One of the ways is to relate it to the housing price and environment in that area. In the following analysis, we design marketing strategy according to the living standard of areas.

```r
PLOT_SIZE_PERC
```


Refrigerator Size Percentage
140T_MORE
120_139T
100_119T
80_99T
60_79T
40_59T
20_39T
LESS_20T
Income Range
0
25
50
75
100
Percentage
RFG_SIZE
Do not use a refrigerator
Half−size or compact
Large (22.6 to 29.5 cubic feet)
Medium (17.6 to 22.5 cubic feet)
Small (less than 17.6 cubic feet)
Very large (more than 29.5 cubic feet)

```
PLOT_SIZE
```



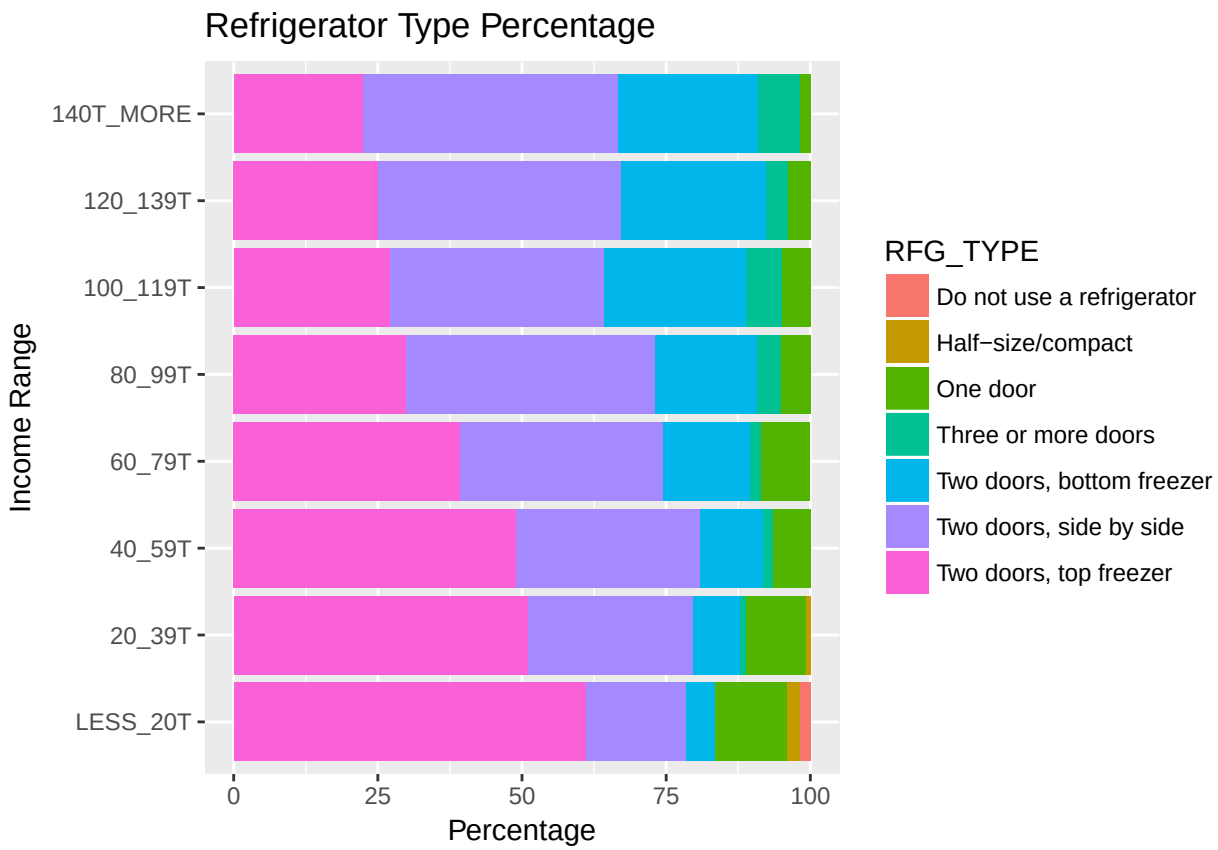
1.distribution of refrigerator sizes among groups of housing units with different income ranges

We first do a percentage plotting and a number of unit plotting for the distribution of refrigerator sizes among different income groups. As we can see from the first plot, with the income increasing, the proportion of housing units using large and very large refrigerators increases while the proportion of housing units using medium and small refrigerators decreases. The proportion of housing units using half-size compact refrigerators and not using refrigerators at all are very small, indicating that they are not the types of refrigerators that refrigerator company should invest on. From the second plot, we can see a similar trend, except that the number of units using large refrigerators doesn't always increase as the income increases. Because the total number of people that are within the high-income range are smaller than those that are within the low-income range, the absolute value of number of units for different types and sizes can't reflect the trend very well. Therefore, we think it makes more sense to design marketing strategy according to the trend in the first plot.

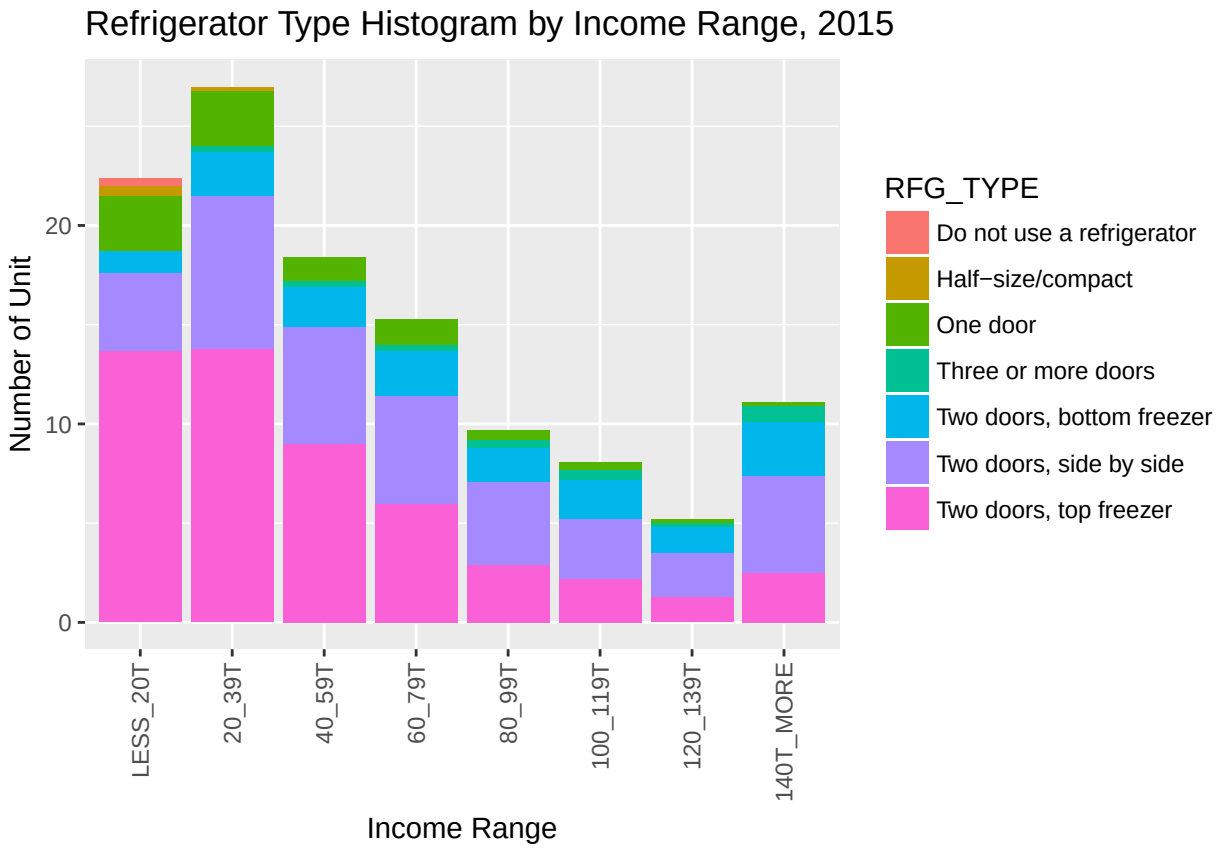
In the areas with high living standard, refrigerators company should focus more on the marketing of large refrigerators. And in very upscale community, company can also make efforts to sell very large refrigerators. But since the proportion of housing units using large refrigerators is much bigger than the one using very large refrigerators, the focus of marketing should be put on large refrigerators in the areas with high living standard.

On the other hand, in the areas with low living standard, refrigerator company should try to sell medium refrigerators. And in very poor areas, they can spend small proportion of budget in marketing small refrigerators. But since the housing units that use medium refrigerators are the main consumers in the areas with low living standard, the refrigerator company should pay attention to advertising and selling medium refrigerators in those areas.

```
PLOT_TYPE_PERC
```


Refrigerator Type Percentage
140T_MORE
120_139T
100_119T
80_99T
60_79T
40_59T
20_39T
LESS_20T
Income Range
0
25
50
75
100
Percentage
RFG_TYPE
Do not use a refrigerator
Half−size/compact
One door
Three or more doors
Two doors, bottom freezer
Two doors, side by side
Two doors, top freezer

```
PLOT_TYPE
```



2.distribution of refrigerator types among groups of housing units with different income ranges

Firstly, the plot shows that the refrigerator types: two doors with top freezer, two doors side by side, two doors with bottom freezer are the most popular ones. As the income increases, the proportion of housing units that use two doors side by side refrigerators and two doors refrigerators with bottom freezer increases while the proportion of housing units that use two doors refrigerators with top freezer decreases. In addition to this, the proportion of housing units that use half-size/compact refrigerators and one door refrigerators are very small.

Based on the plot, we can say that in the areas with high living standard (income higher than 80,000), the emphasis of marketing should fall on two door side by side refrigerators and two doors refrigerators with bottom freezer. And in more upscale areas, more focus should be put on two doors refrigerators with bottom freezer. But the majority of marketing budget should always be spent on the two-door side by side refrigerators since they are the most popular ones.

In the areas with middle living standard (income between 60,000 to 80,000), the refrigerator company should pay more attention to the marketing of two door refrigerators with top freezer. The type preference is similar in the areas with low living standard (income lower than 60,000) except that company could consider selling one door refrigerators in the areas with extremely low living standard.

In our analysis, we mostly use the percentage plotting to see which kind of refrigerators we should care about during marketing. But number of unit plotting also gives us valuable information that housing units from middle and lower classes are the main consumers of refrigerators. Therefore, refrigerator companies should focus on marketing refrigerators in the areas with middle and low living standard.

Conclusion

In general, medium size and two doors refrigerators with top freezers are the most popular ones throughout U.S, especially in the West North Central and East North Central. Therefore, they can be the biggest target for marketing in the whole U.S. And the large size refrigerators can be another target for marketing, since they are more popular than the medium size refrigerators in the Middle Atlantic, New England, West Central and South Atlantic.

Viewing from the living standard perspective, refrigerator companies should focus on marketing large refrigerators with two doors side by side and two doors with bottom freezer in the areas with high living standard. In the areas with middle and lower living standard, refrigerator companies could spend more time and money on marketing two doors medium refrigerators with top freezer.

Because our analysis only analyzes the data of different income range and regions, we don't have a complete picture of the distribution of refrigerators features among consumers. In the future, we hope we can explore more data and gain more useful suggestions on marketing for refrigerator companies.