Portforlio#5

Ruinan(Victor) Zhang

10/12/2016

# Background

Given the median salary of locations in United States and using web scraping tools in R such as "Rvest", this task requires a statistical analysis based on data to determine if Target should open a store in Terre Haute area. The data of median house income is provided by instructor and the number of Target stores at locations are scrabed from online.

# Approach

The ensential skills required for accomplishing this task is to properly scrape information from website and cleaning data for usage. R Packages such as "Rvest" and "stringr" are used to do web scraping and data cleaning. The Target store infomation of each state is obtained in every iterations of the for loop. Inside the for loop, the state's corresponding url is created by combining two base url components and state name. Then, after identifying the html class containing desired informations, functions form "rvest" are called to extract information as in pure text. String manipulations are applied to the text to clean into a nice dataframe which can be used later.

## Loading required package: xml2

##   
## Attaching package: 'rvest'

## The following object is masked from 'package:readr':  
##   
## guess\_encoding

# set base urls  
base\_url1 <- "http://www.allstays.com/c/target-"  
base\_url2 <- "-locations.htm"  
  
df\_stores <- data.frame() # dataframe containing information of zipcode where is at least a Target Store  
  
# iterate over states' abbreviations  
for (state\_brev in state.abb){  
 state<-state.name[grep(state\_brev,state.abb)] # get state name based on abbreviation  
 state <- tolower(state) # change state name to lower case  
 state <- sub(" ","-",state) # add "-" to state name  
 url <- paste(base\_url1,state,base\_url2,sep = "") # construct URL  
 html<- read\_html(url) # read html  
   
 txt <-html\_nodes(html,".col-md-5") %>% html\_text() # extract texts from specific html class  
 all\_zips <- as.vector(str\_extract\_all(txt,paste(state\_brev,"[0-9][0-9][0-9][0-9][0-9]"))) #extract zipcode  
 to\_remove = paste(state\_brev,"")  
 all\_zips <- gsub(to\_remove,"",unlist(all\_zips)) # remove state abbreviation in the zipcode  
 df\_state <- as.data.frame(table(all\_zips)) # construct datafrome for zipcodes which has Target Store  
 df\_stores<- rbind(df\_stores,df\_state) # attach result to final output df  
}  
names(df\_stores) <- c("ZipCode","#stores") # rename column names  
save(df\_stores,file = "target\_info.Rda") # save result

After obtaing alll informations of targets at 50 states of united states, the result is merged with the informations of median income by zipcode.

income\_df = read.csv("input/ACS\_14\_5YR\_B19013\_with\_ann.csv",skip =1)  
# rename some columns  
colnames(income\_df)[4]<-"estimate\_median"  
colnames(income\_df)[5]<-"marginal\_error"  
  
income\_df$Geography=gsub("ZCTA5 ","" ,income\_df$Geography) #cleaning zipcode column  
income\_df$Geography = as.factor(income\_df$Geography) # change data types to factor  
income\_df$estimate\_median = as.numeric(income\_df$estimate\_median) # change data types to numerical  
  
df\_income <- income\_df[c("Geography","estimate\_median")]  
names(df\_income) <- c("ZipCode","est\_income\_median")  
  
load("target\_info.Rda") # import Target store information  
df\_gen <- merge(df\_income,df\_stores,by="ZipCode",all = TRUE)

# Result

After acquired all necessary information, the criteria for determination of whether Target should open a store in Terre Haute area is to compare number of stores where median income exceeds Terre Haute area and the number of stores where median income does not exceed Terre Haute.

TH\_median <- df\_gen$est\_income\_median[df\_gen$ZipCode==47803] # median income of Terre Haute  
  
# number of Target where median is lower higher  
store\_higher\_median <- sum(df\_gen$est\_income\_median>TH\_median & df\_gen$`#stores`>0,na.rm=TRUE)  
# number of Target where median income is lower  
store\_lower\_median <- sum(df\_gen$est\_income\_median<TH\_median & df\_gen$`#stores`>0,na.rm=TRUE)

Number of Target stores where median income exceeds Terre Haute:

## [1] 1150

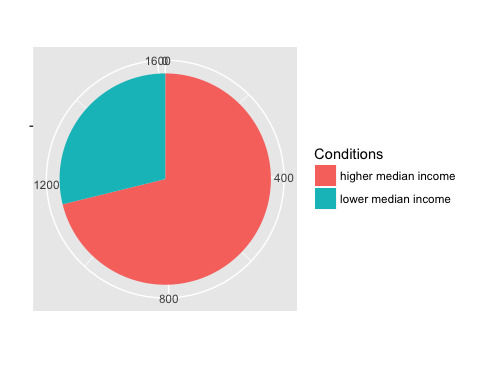
Number of Target stores where median income does not exceed Terre Haute:

## [1] 466

Considering there are over 400 current openning Target stores are located where median income is lower than Terre Haute area, if median income is the only factor for the decision making, it will probably be profitable for Target to open a store in Terre Haute area.

#### Graph

Based on the number of locations where has targets, a pie plot is generated to have a better visualizaiton of ratio between the locations where median income is higher or lower than Terre Haute.



Looking at the pie plot, there are roughly over 30% of targets are opened in the area where median income is lower than Terre Haute. Based on a significant amount of Targets open in those area with lower incomes, one should be open in Terre Haute.

# Reflection

Accomplishing this project requires basic knowledge of html structure and rvest library. The most time-consuming part of this project is to identify the desired html node from viewing the webpage source code. In order to filting the zipcode, I learned regular expression to perform the string matching function in "stringr" library.

#### Limitations

Idealy, the number of locations where do not have Target should be taken into consideration as well. However, even though the income information of all location is provided in the dataset, information on the population and other geographical information are not provided. Without those necessary information, the fact that many locations with low population do not have targets and they can make up majority of the dataset can raise a strong bias. Therefore, an inspection of numbers of locations without Target is not conducted based on limited recourses.

# Full Code

library(readr)  
library(rvest)  
library(jsonlite)  
library(httr)  
library(xml2)  
library(stringr)  
  
# set base urls  
base\_url1 <- "http://www.allstays.com/c/target-"  
base\_url2 <- "-locations.htm"  
  
df\_stores <- data.frame() # dataframe containing information of zipcode where is at least a Target Store  
  
# iterate over states' abbreviations  
for (state\_brev in state.abb){  
 state<-state.name[grep(state\_brev,state.abb)] # get state name based on abbreviation  
 state <- tolower(state) # change state name to lower case  
 state <- sub(" ","-",state) # add "-" to state name  
 url <- paste(base\_url1,state,base\_url2,sep = "") # construct URL  
 html<- read\_html(url) # read html  
   
 txt <-html\_nodes(html,".col-md-5") %>% html\_text() # extract texts from specific html class  
 all\_zips <- as.vector(str\_extract\_all(txt,paste(state\_brev,"[0-9][0-9][0-9][0-9][0-9]"))) #extract zipcode  
 to\_remove = paste(state\_brev,"")  
 all\_zips <- gsub(to\_remove,"",unlist(all\_zips)) # remove state abbreviation in the zipcode  
 df\_state <- as.data.frame(table(all\_zips)) # construct datafrome for zipcodes which has Target Store  
 df\_stores<- rbind(df\_stores,df\_state) # attach result to final output df  
}  
names(df\_stores) <- c("ZipCode","#stores") # rename column names  
save(df\_stores,file = "target\_info.Rda") # save result  
  
  
income\_df = read.csv("input/ACS\_14\_5YR\_B19013\_with\_ann.csv",skip =1)  
# rename some columns  
colnames(income\_df)[4]<-"estimate\_median"  
colnames(income\_df)[5]<-"marginal\_error"  
  
income\_df$Geography=gsub("ZCTA5 ","" ,income\_df$Geography) #cleaning zipcode column  
income\_df$Geography = as.factor(income\_df$Geography) # change data types to factor  
income\_df$estimate\_median = as.numeric(income\_df$estimate\_median) # change data types to numerical  
  
df\_income <- income\_df[c("Geography","estimate\_median")]  
names(df\_income) <- c("ZipCode","est\_income\_median")  
  
load("target\_info.Rda") # import Target store information  
df\_gen <- merge(df\_income,df\_stores,by="ZipCode",all = TRUE)  
  
TH\_median <- df\_gen$est\_income\_median[df\_gen$ZipCode==47803] # median income of Terre Haute  
  
  
# number of Target where median is lower higher  
store\_higher\_median <- sum(df\_gen$est\_income\_median>TH\_median & df\_gen$`#stores`>0,na.rm=TRUE)  
# number of Target where median income is lower  
store\_lower\_median <- sum(df\_gen$est\_income\_median<TH\_median & df\_gen$`#stores`>0,na.rm=TRUE)  
  
groups <- c("higher median income","lower median income")  
info <- rbind(store\_higher\_median,store\_lower\_median)  
df\_tem <- data.frame(groups,info) # create data frame for plot usage  
  
g <- ggplot(df\_tem,aes(x="",y=info,fill=groups))+geom\_bar(width = 1,stat = "identity") # construct bar plot  
g <- g+guides(fill=guide\_legend(title="Conditions"))+xlab("")+ylab("") # set labels  
g <- g+coord\_polar("y",start=0) # change to polar coordinate system  
g