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 an 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 are scrabed from online.

# Approach

The ensential skills required for complishing this task is to properly scrape information from website and cleaning data for usage. R Packages such as "Rvest" and "stringr" are used to accomplish the goal. The Target store infomation of one state is obtained in every iterations of the for loop. The state's corresponding url is created by two base url components and state name. Then after identity the html class containing desired informations, functions form "rvest" are called to extract pure text informations. Some string manipulations are applied to the text to clean the text into a nice dataframe which can be used later. After obtaing alll informations of targets at 50 states of united states, the result is merged with the informations of median income by zipcode.

# 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)