# praca domowa: pobranie danych z innego portalu (np. otomoto)

pobieranie dla różnych marek z listy, a dla marek gdzie podstron jest więcej niż 500 to wyszukiwanie odbywa się dla różnych lat

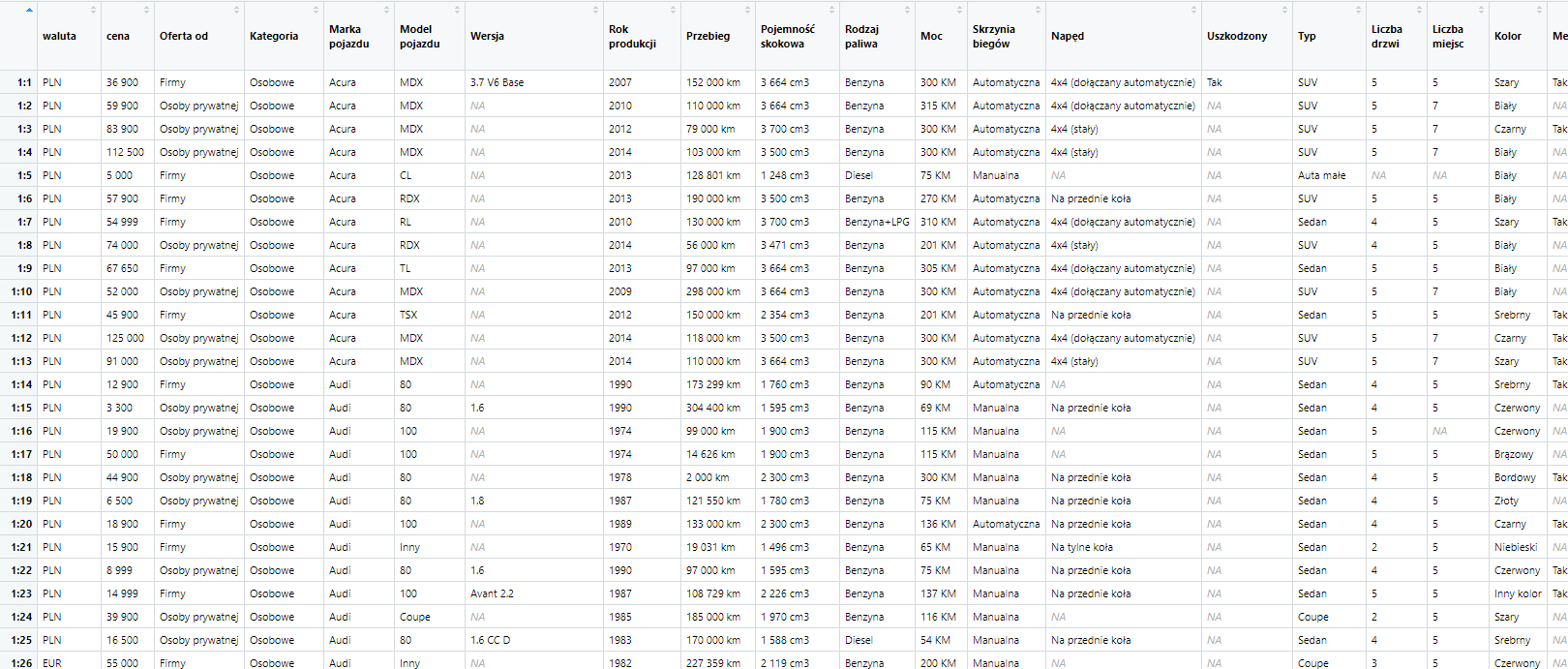
rozwiązania:

* wersja rozszerzona w pliku pd2\_scrap\_v2.R
* wersja podstawowa w pliku pd2\_scrap.R

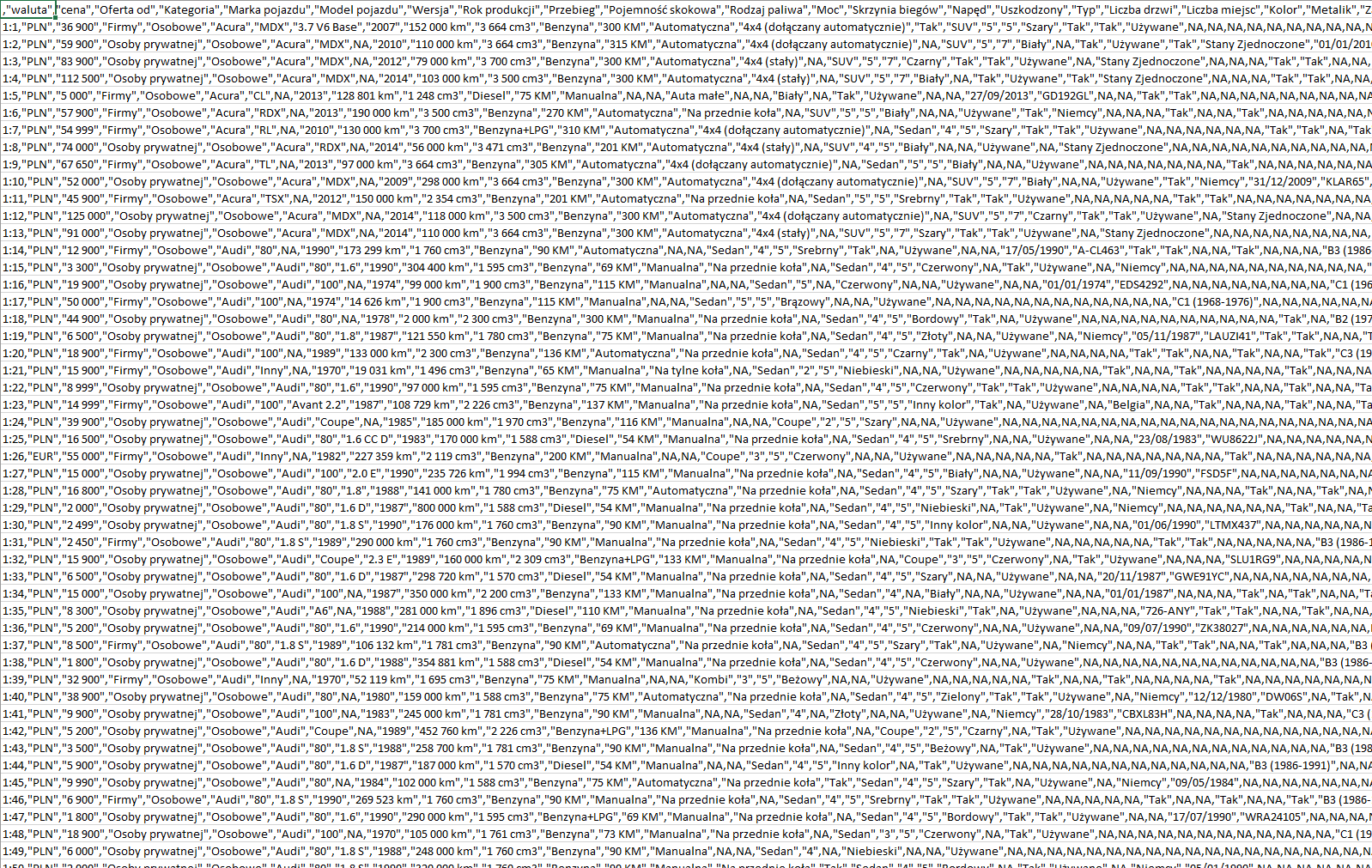
## Zrzut ekranu z pobierania linku dla różnych marek i lat



## Przykładowe wyniki:



## Z pliku:



## Kod wersja zawansowana - pobieranie dla roznych marek z listy, a dla marek gdzie podstron jest wiecej niz 5000 to wyszukiwanie odbywa sie dla róznych lat

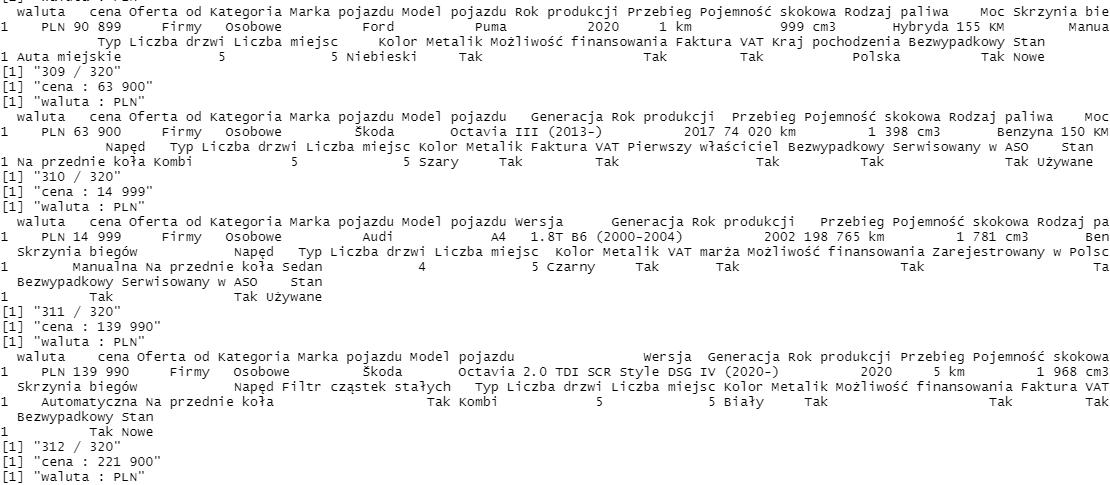
|  |
| --- |
| # praca domowa: pobranie danych z innego portalu (np. otomoto)  # pobieranie dla roznych marek z listy, a dla marek gdzie podstron jest wiecej niz 5000  # to wyszukiwanie odbywa sie dla róznych lat  library(seleniumPipes)  library(dplyr)  library(stringr)  library(gtools)  library(rvest)  marki <- c('Acura', 'audi')  #marki <- c('Acura', 'Aixam', 'Alfa-Romeo', 'Aston-Martin', 'audi', 'austin', 'autobianchi')  #'volkswagen', 'bmw', 'opel', 'ford', 'mercedes-benz','renault', 'skoda', 'toyota')  i <-1  wektorLinkowOtoMotoU<-c()  for(marka in marki){  print(marka)  newUrl2<- paste0("https://www.otomoto.pl/osobowe/", marka, "/?page=",i)  page<-read\_html(newUrl2)  licznikAut <- page %>% html\_nodes(xpath = "//span[@class='fleft tab selected']/span[2]") %>% html\_text()  licznikAut <- gsub('[()]+','',licznikAut)  licznikAut <- as.integer(str\_replace(licznikAut, " ", ""))  licznikAut  print(paste0("licznikAut: ",licznikAut))  ileStron = ceiling(licznikAut/32)  print(paste0("ileStron: ",ileStron))    if(ileStron > 500)  {  i <- 1  for(rok in 1990: 2021)  {  if(rok == 1990){  newUrl3<- paste0("https://www.otomoto.pl/osobowe/", marka, "/?search[filter\_float\_year%3Ato]=",rok )  print(paste0("wszystkie egzempalrze do ", rok, " dla marki ", marka))  }  else{  newUrl3<- paste0("https://www.otomoto.pl/osobowe/", marka, "/od-", rok, "/?search[filter\_float\_year%3Ato]=",rok)  print(paste0("rok ", rok, " dla marki ", marka))  }  print(newUrl3)  page<-read\_html(newUrl3)  licznikAut <- page %>% html\_nodes(xpath = "//span[@class='fleft tab selected']/span[2]") %>% html\_text()  licznikAut <- gsub('[()]+','',licznikAut)  licznikAut <- as.integer(str\_replace(licznikAut, " ", ""))  print(paste0("licznikAut: ",licznikAut))  if(licznikAut > 0)  {  ileStron = ceiling(licznikAut/32)  print(paste0("ileStron: ",ileStron))  wektorLinkowOtoMoto<-c()  for(i in 1:ileStron){  print(paste0(i," z ",ileStron))  if(rok == 1990){  newUrl<- paste0("https://www.otomoto.pl/osobowe/", marka, "/?search[filter\_float\_year%3Ato]=",rok,"&page=",i )  }  else{  newUrl<- paste0("https://www.otomoto.pl/osobowe/", marka, "/od-", rok, "/?search[filter\_float\_year%3Ato]=",rok,"&page=",i )  }  page<-read\_html(newUrl)  result<-page%>%html\_nodes(xpath='/html/body/div[4]/div[2]/section/div[2]/div[1]/div/div[1]/div[6]/article[\*]/div[2]/div[1]/div[1]/h2/a')  #print(result)  wektorLinkowOtoMoto<-c(wektorLinkowOtoMoto,xml\_attr(result,"href"))  #print(wektorLinkowOtoMoto)  }  wektorLinkowOtoMotoU\_temp<-wektorLinkowOtoMoto%>%unique()  wektorLinkowOtoMotoU <- c(wektorLinkowOtoMotoU, wektorLinkowOtoMotoU\_temp)  }  }  }  else  {  ileStron = ceiling(licznikAut/32)  wektorLinkowOtoMoto<-c()  for(i in 1:ileStron){  print(paste0(i," z ",ileStron))  newUrl<- paste0("https://www.otomoto.pl/osobowe/", marka, "/?page=",i)  page<-read\_html(newUrl)  result<-page%>%html\_nodes(xpath='/html/body/div[4]/div[2]/section/div[2]/div[1]/div/div[1]/div[6]/article[\*]/div[2]/div[1]/div[1]/h2/a')  #print(result)  wektorLinkowOtoMoto<-c(wektorLinkowOtoMoto,xml\_attr(result,"href"))  #print(wektorLinkowOtoMoto)  }  wektorLinkowOtoMotoU\_temp<-wektorLinkowOtoMoto%>%unique()  wektorLinkowOtoMotoU <- c(wektorLinkowOtoMotoU, wektorLinkowOtoMotoU\_temp)  }  }  print(wektorLinkowOtoMotoU)  zrobWierszRvestOtoMoto<-function(w,wektorLinkow,remDr){  newUrl<-wektorLinkow[w]  page<-read\_html(newUrl)  cena<-html\_node(page,".offer-price\_\_number")%>%html\_text()  cena <- gsub('[A-Z]+','',cena)  cena <- str\_trim(cena)  cena <- str\_replace\_all(cena, "[\r\n]" , "")  #print(paste0("cena : ",cena ) )  waluta<-html\_node(page,".offer-price\_\_currency")%>%html\_text()  #print(paste0("waluta : ",waluta ) )  v1 <-html\_nodes(page, ".offer-params\_\_label")%>%html\_text()%>%na.omit()  v1 <- str\_replace\_all(v1, "[\r\n]" , "")  v2 <-html\_nodes(page, ".offer-params\_\_value")%>%html\_text()%>%na.omit()  v2 <- str\_replace\_all(v2, "[\r\n]" , "")  v2 <- gsub(' +',' ',v2)  v2 <- str\_trim(v2)  nazwyKolumn <- v1 #[indexy%%2==1]  #print('nazwyKolumn')  #print(nazwyKolumn)  wartosci <- v2 #[indexy%%2==0]  #print('wartosci')  #print(wartosci)  df1<- data.frame (matrix(wartosci,nrow = 1,ncol=length(wartosci)) )  names(df1) <- nazwyKolumn  df1<-cbind(cena,df1)  df1<-cbind(waluta,df1)  #print(df1)  }  mieszkania3OtoMoto<-NULL  liczbaLinkowOtoMoto<-length(wektorLinkowOtoMotoU)  #for(w in 1: 3 ){  for(w in 1: liczbaLinkowOtoMoto ){  print(paste0(w," / ",liczbaLinkowOtoMoto ) )  skip<-FALSE  tryCatch(  df1<-zrobWierszRvestOtoMoto(w,wektorLinkowOtoMotoU,remDr),error=function(e){skip<<-TRUE}  )  # print(df1)  if(skip){next}  if(is.null(mieszkania3OtoMoto)){  mieszkania3OtoMoto<-df1  }else{  mieszkania3OtoMoto<-smartbind(mieszkania3OtoMoto,df1)  }  }  mieszkania3OtoMoto  write.csv(mieszkania3OtoMoto, "mieszkania3OtoMoto.csv" ) |

# Wersja uproszczona

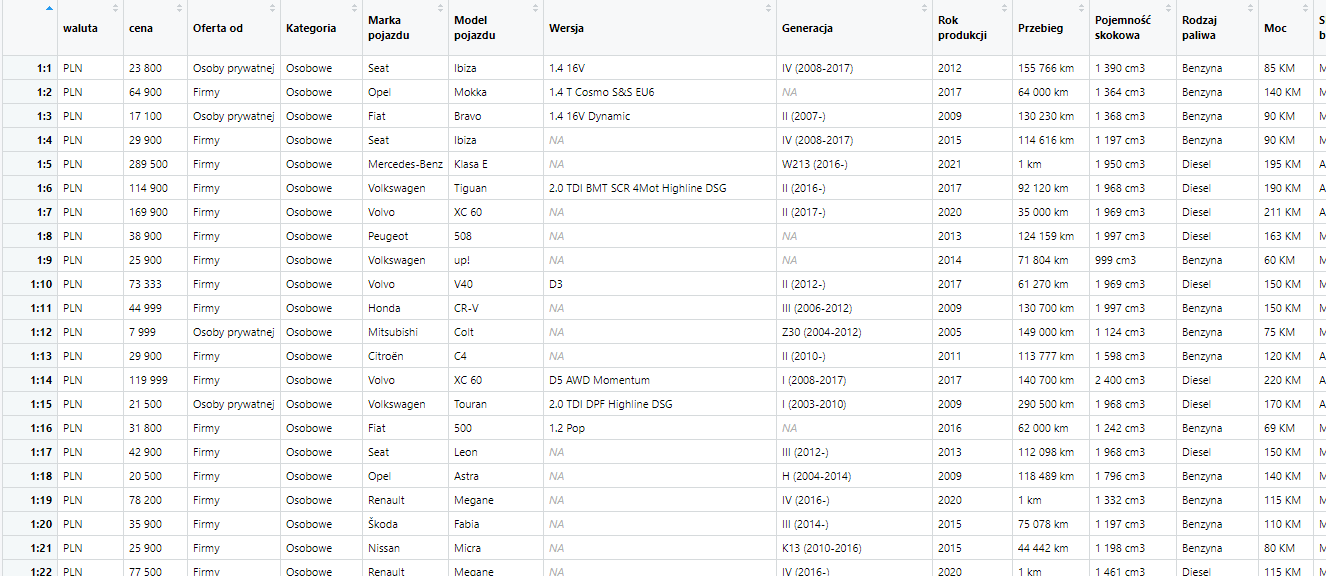
Zrzut ekranu z pobierania adresów strony



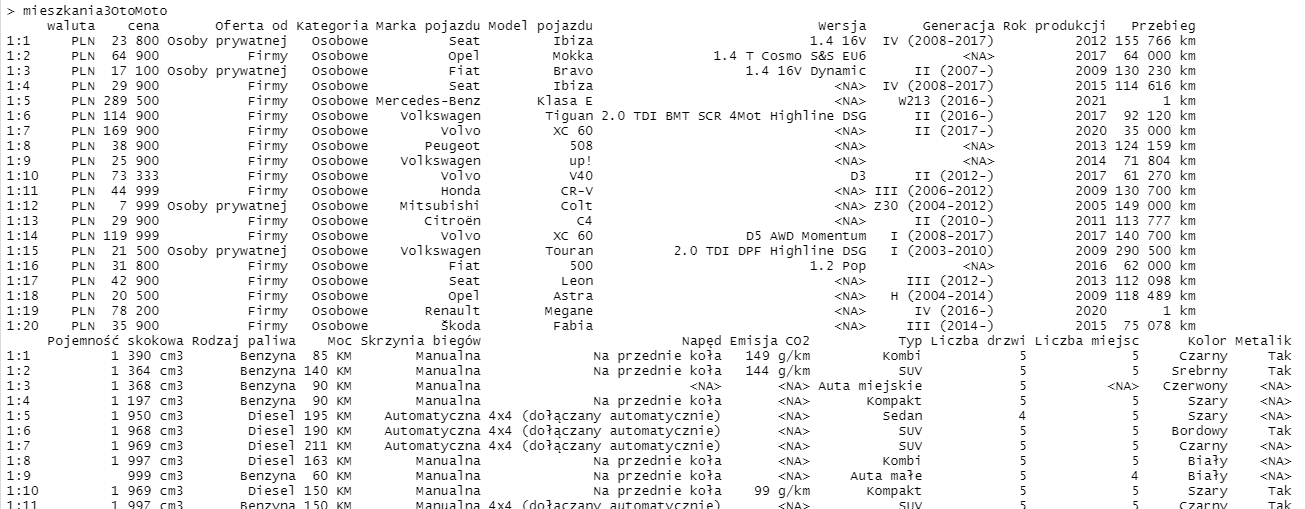
Rzut ekranu fragmentu działania (pobierania danych ze wskazanej strony)



Zrzut ekranu z tabeli wynikowej



Zrzut ekranu z konsoli wyświetlającej tabele



Kod wersja uproszczona

|  |
| --- |
| # praca domowa: pobranie danych z innego portalu (np. otomoto)  library(seleniumPipes)  library(dplyr)  library(stringr)  library(gtools)  library(rvest)  wektorLinkowOtoMoto<-c()  for(i in 1:10){  newUrl<- paste0("https://www.otomoto.pl/osobowe/?page=",i)  page<-read\_html(newUrl)  result<-page%>%html\_nodes(xpath='/html/body/div[4]/div[2]/section/div[2]/div[1]/div/div[1]/div[6]/article[\*]/div[2]/div[1]/div[1]/h2/a')  print(result)  wektorLinkowOtoMoto<-c(wektorLinkowOtoMoto,xml\_attr(result,"href"))  print(wektorLinkowOtoMoto)  }  wektorLinkowOtoMotoU<-wektorLinkowOtoMoto%>%unique()  zrobWierszRvestOtoMoto<-function(w,wektorLinkow,remDr){  newUrl<-wektorLinkow[w]  page<-read\_html(newUrl)  cena<-html\_node(page,".offer-price\_\_number")%>%html\_text()  cena <- gsub('[A-Z]+','',cena)  cena <- str\_trim(cena)  cena <- str\_replace\_all(cena, "[\r\n]" , "")  print(paste0("cena : ",cena ) )  waluta<-html\_node(page,".offer-price\_\_currency")%>%html\_text()  print(paste0("waluta : ",waluta ) )  v1 <-html\_nodes(page, ".offer-params\_\_label")%>%html\_text()%>%na.omit()  v1 <- str\_replace\_all(v1, "[\r\n]" , "")  v2 <-html\_nodes(page, ".offer-params\_\_value")%>%html\_text()%>%na.omit()  v2 <- str\_replace\_all(v2, "[\r\n]" , "")  v2 <- gsub(' +',' ',v2)  v2 <- str\_trim(v2)  nazwyKolumn <- v1 #[indexy%%2==1]  #print('nazwyKolumn')  #print(nazwyKolumn)  wartosci <- v2 #[indexy%%2==0]  #print('wartosci')  #print(wartosci)  df1<- data.frame (matrix(wartosci,nrow = 1,ncol=length(wartosci)) )  names(df1) <- nazwyKolumn  df1<-cbind(cena,df1)  df1<-cbind(waluta,df1)  print(df1)  }  mieszkania3OtoMoto<-NULL  liczbaLinkowOtoMoto<-length(wektorLinkowOtoMotoU)  #for(w in 1: 3 ){  for(w in 1: liczbaLinkowOtoMoto ){  print(paste0(w," / ",liczbaLinkowOtoMoto ) )  skip<-FALSE  tryCatch(  df1<-zrobWierszRvestOtoMoto(w,wektorLinkowOtoMotoU,remDr),error=function(e){skip<<-TRUE}  )  # print(df1)  if(skip){next}  if(is.null(mieszkania3OtoMoto)){  mieszkania3OtoMoto<-df1  }else{  mieszkania3OtoMoto<-smartbind(mieszkania3OtoMoto,df1)  }  }  mieszkania3OtoMoto |