**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT**



**KHOA ĐIỆN – ĐIỆN TỬ**

**----------🙞🙜----------**



**BÀI TẬP CHUYÊN ĐỀ CƠ SỞ DỮ LIỆU**

**Đề: scrapping data from a website**

**SVTH : Võ Thành Luân**

**MSV : 1811505310324**

**Lớp : 220CCSDL03**

**GVHD : ThS. Nguyễn Thị Thúy Hoài**

*Đà Nẵng, tháng 05 năm 2021*

**Mục lục**

[1. Sơ đồ ERD 1](#_Toc76493192)

[2. Các bước thực hiện 1](#_Toc76493193)

[2.1. Cài đặt các thư viện 1](#_Toc76493194)

[2.2. Load các thư viện 1](#_Toc76493195)

[2.3. Gán link website cần lấy dữ liệu 2](#_Toc76493196)

[2.4. Lấy dữ liệu cho bảng thương hiệu 2](#_Toc76493197)

[2.4.1. Code lấy dữ liệu 2](#_Toc76493198)

[2.4.2. Cách lấy node 2](#_Toc76493199)

[2.4.3. Kết quả dữ liệu 3](#_Toc76493200)

[2.5. Lấy dữ liệu và đường dẫn của từng loại sản phẩm 4](#_Toc76493201)

[2.5.1. Code lấy dữ liệu 4](#_Toc76493202)

[2.5.2. Cách lấy các node 5](#_Toc76493203)

[2.5.3. Kết quả dữ liệu 5](#_Toc76493204)

[2.6. Lấy dữ liệu của sản phẩm dựa vào đường dẫn của từng loại 6](#_Toc76493205)

[2.6.1. Code lấy dữ liệu 6](#_Toc76493206)

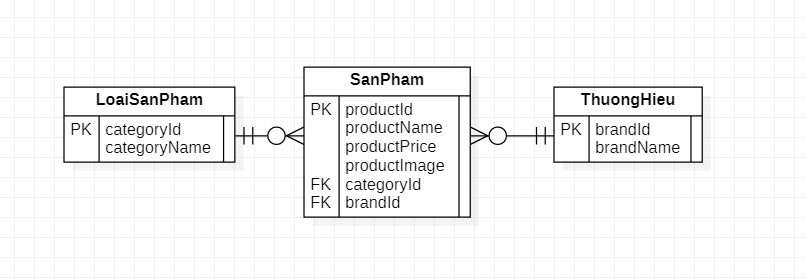
[2.6.2. Lấy các node và thông tin của sản phẩm theo từng loại 8](#_Toc76493207)

[2.6.3. Tạo thông tin khóa ngoại giữa 2 bảng loại sản phẩm và sản phẩm 8](#_Toc76493208)

[2.6.4. Xuất 3 bảng dữ liệu vào 3 sheet khác nhau 8](#_Toc76493209)

[2.6.5. Kết quả dữ liệu 9](#_Toc76493210)

1. Sơ đồ ERD



1. Các bước thực hiện
   1. Cài đặt các thư viện

install.packages('xml2')

install.packages('rvest')

install.packages("writexl")

install.packages('selectr')

install.packages("xlsx")

install.packages("magrittr")

install.packages("dplyr")

install.packages("tidyverse")

install.packages("rebus")

install.packages("lubridate")

* 1. Load các thư viện

library(xml2)

library(rvest)

library(writexl)

library(stringr)

library(xlsx)

library(magrittr)

library(dplyr)

library(tidyverse)

library(rebus)

library(lubridate)

* 1. Gán link website cần lấy dữ liệu

url <- 'https://tainghe.com.vn/tai-nghe.html'

homepage <- read\_html(url)

* 1. Lấy dữ liệu cho bảng thương hiệu
     1. Code lấy dữ liệu

**#lay du lieu cho bang thuong hieu**

**#lay danh sach cac thuong hieu**

brand\_html <- html\_nodes(homepage, xpath="/html/body/div[3]/div[5]/div[1]/div[4]/div[1]/ul[2]/li/a")

brands <- html\_text(brand\_html)

brandName <-c()

brandId <-c()

ThuongHieu <- data.frame(brandId, brandName)

i<-1

for (name in brands){

**#lay ten thuong hieu**

brandName <-c(name)

**#lay ma thuong hieu**

brandId <- c(str\_c("BRAND",i))

temp <- data.frame(brandId, brandName)

ThuongHieu <- rbind(ThuongHieu,temp)

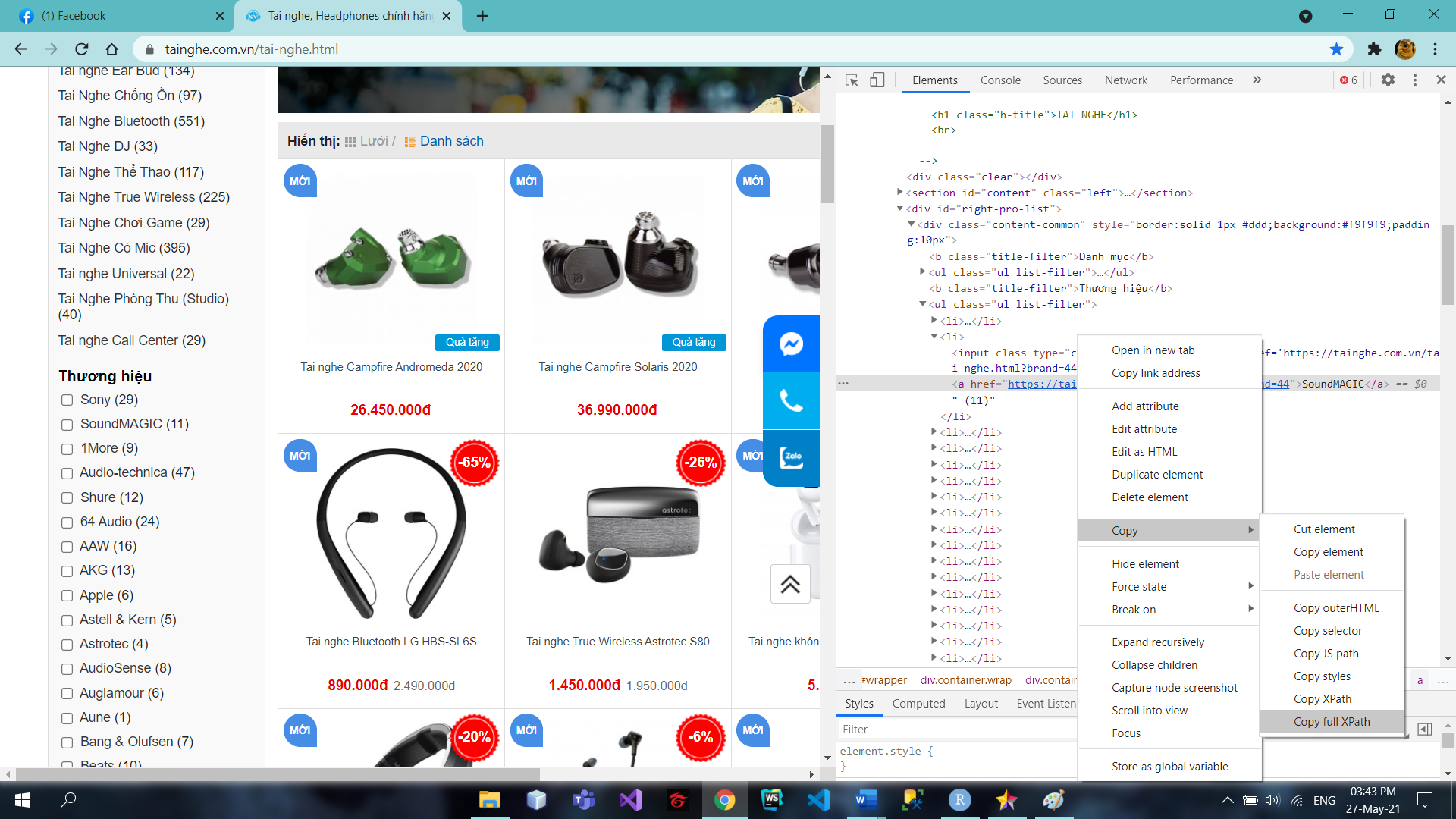
i<-i+1

}

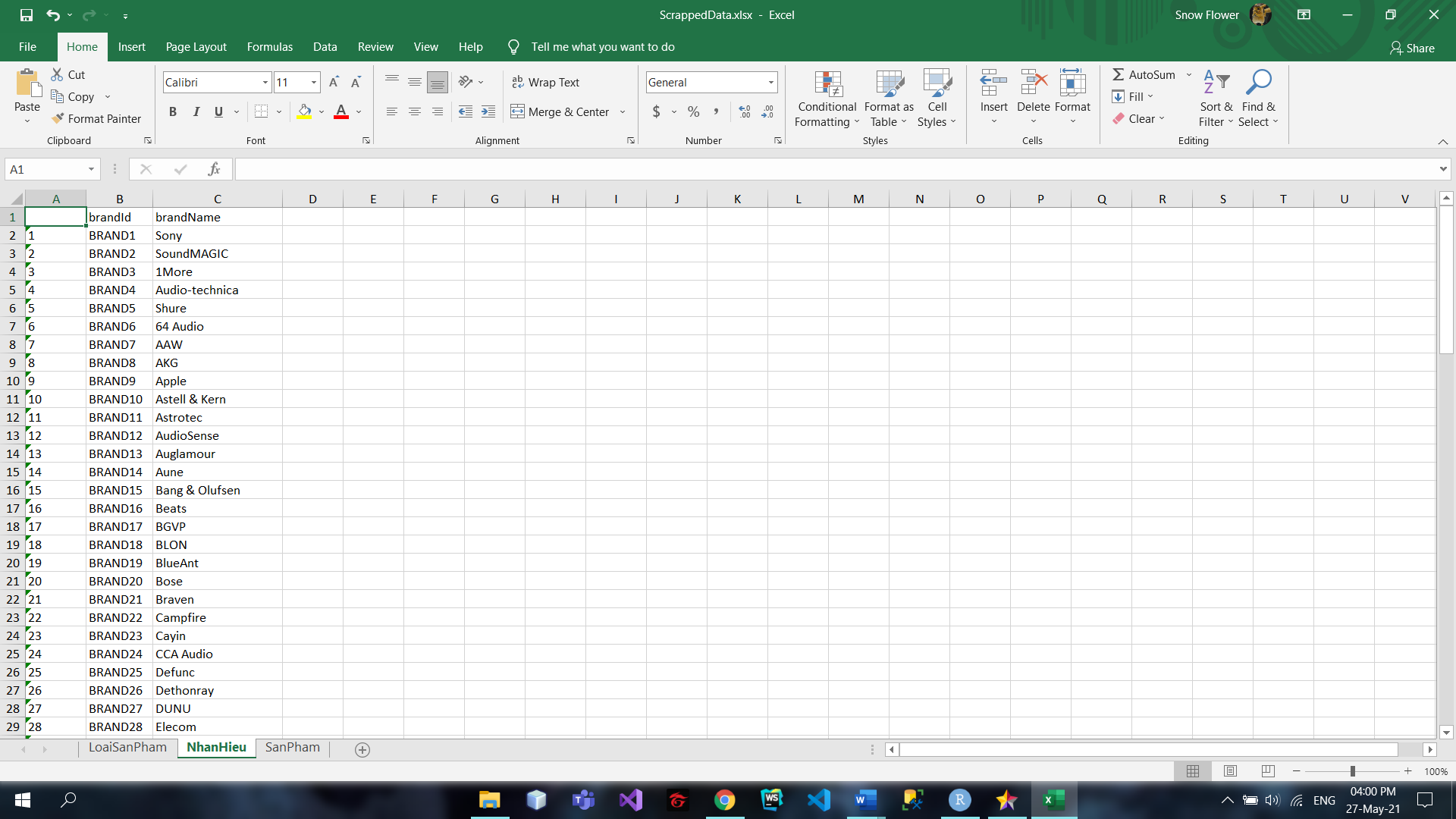
* + 1. Cách lấy node

Lấy node xpath của thương hiệu bằng cách copy full xpath cần lấy:  
**/html/body/div[3]/div[5]/div[1]/div[4]/div[1]/ul[2]/li[2]/a**

Bỏ đi **[2]** trong **li** để lấy toàn bộ node danh sách thương hiệu



* + 1. Kết quả dữ liệu



* 1. Lấy dữ liệu và đường dẫn của từng loại sản phẩm
     1. Code lấy dữ liệu

**#lay du lieu cho bang loai san pham**

**#lay danh sach cac loai san pham**

category\_html<-html\_nodes(homepage, xpath="/html/body/div[3]/div[5]/div[1]/div[4]/div[1]/ul[1]/li/a")

categories\_url<-html\_attr(category\_html,"href")

categories\_url<-str\_c("https://tainghe.com.vn",categories\_url)

category\_name<-html\_text(category\_html)

category\_name<-str\_trim(category\_name)

categoryId <-c()

categoryName <-c()

categoryHref <-c()

LoaiSanPham <- data.frame(categoryId, categoryName, categoryHref)

i<-1

for (name in category\_name){

**#lay ma loai san pham**

categoryId <- c(str\_c("TYPE",i))

**#lay ten loai san pham**

categoryName <-c(name)

**#lay duong dan loai san pham**

categoryHref <-c(categories\_url[i])

temp <- data.frame(categoryId, categoryName, categoryHref)

LoaiSanPham <- rbind(LoaiSanPham,temp)

i<-i+1

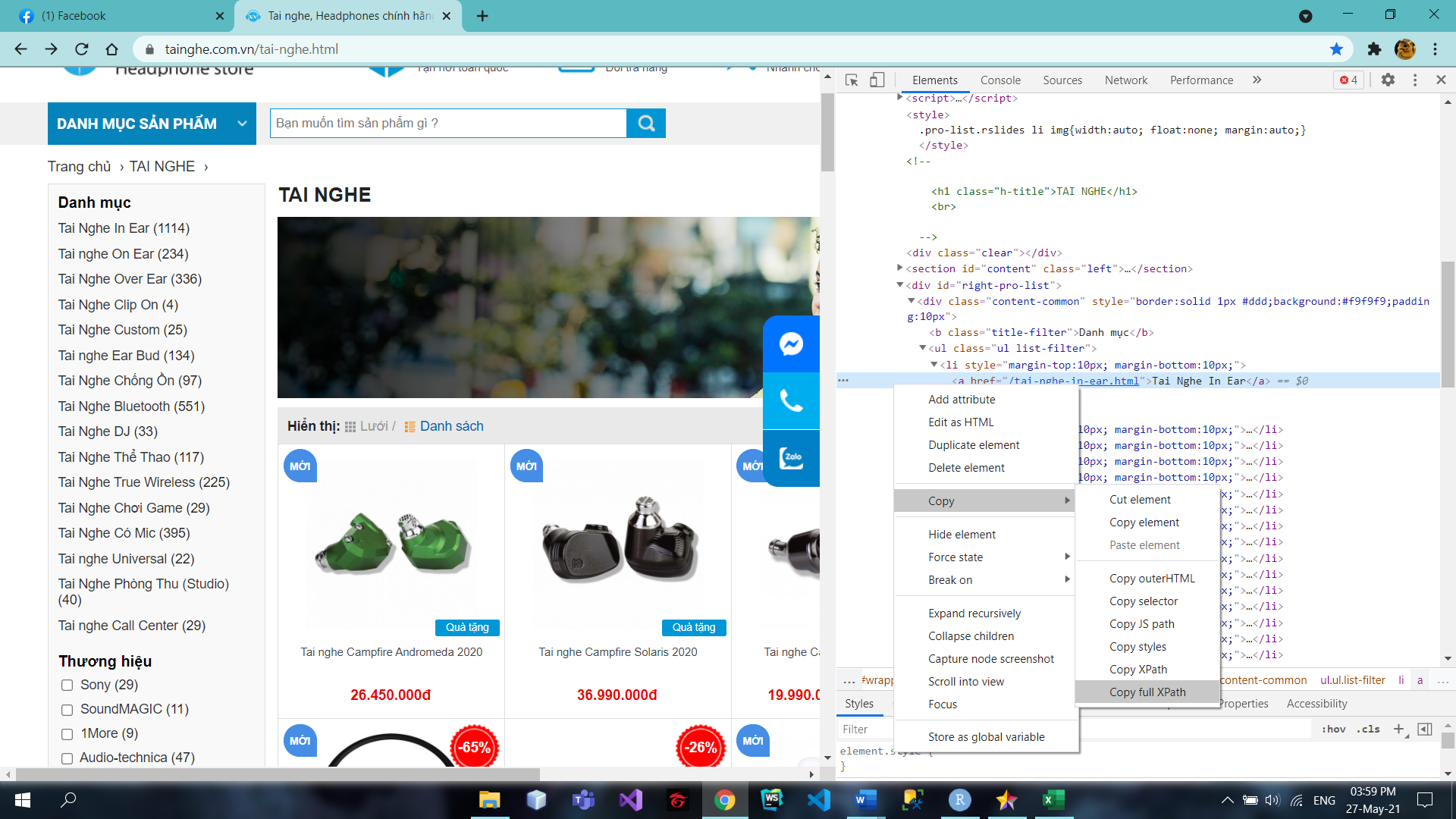
}

* + 1. Cách lấy các node

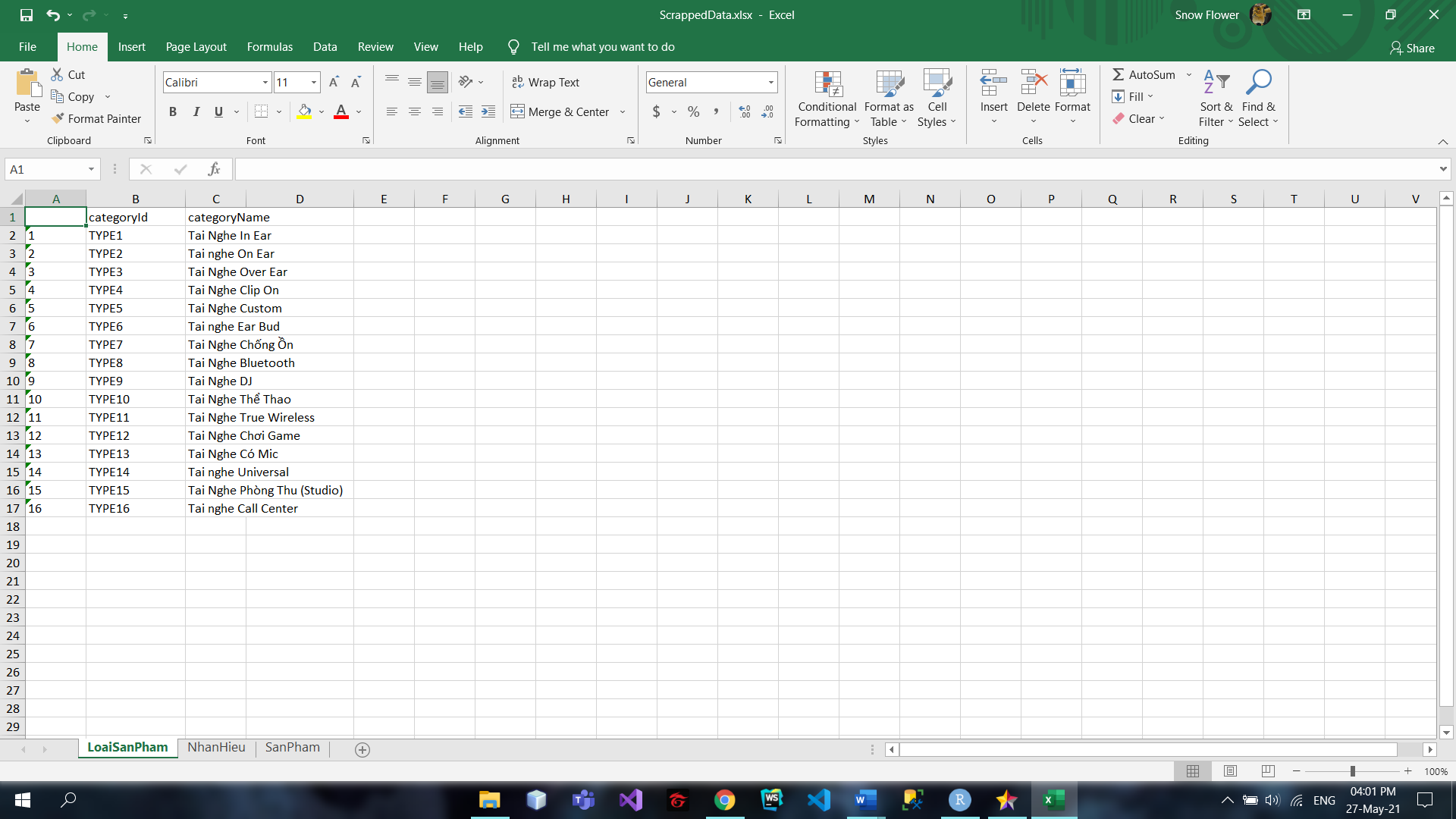
Lấy node xpath của loại bằng cách copy full xpath cần lấy:

**/html/body/div[3]/div[5]/div[1]/div[4]/div[1]/ul[1]/li[1]/a**

Loại bỏ **[1]** trong thẻ **li** để lấy toàn bộ loại sản phẩm



* + 1. Kết quả dữ liệu



* 1. Lấy dữ liệu của sản phẩm dựa vào đường dẫn của từng loại
     1. Code lấy dữ liệu

productId <- c()

productName <- c()

productPrice <- c()

productImage <- c()

categoryHref <-c()

SanPham <- data.frame(productId, productName, productPrice, productImage, categoryHref)

i<-1

for(category\_url in categories\_url){

if(str\_starts(category\_url,"https://tainghe.com.vn/")== FALSE){

category\_url<- str\_c("https://tainghe.com.vn",category\_url)

}

if(str\_starts(category\_url,"https://tainghe.com.vn")){

category\_page<-read\_html(category\_url)

products\_html<-html\_nodes(category\_page,".item div")

products\_detail\_url<-html\_nodes(products\_html,"a")

products\_detail\_url<-html\_attr(products\_detail\_url,"href")

products\_detail\_url<- str\_c("https://tainghe.com.vn",products\_detail\_url)

for(product\_detail\_url in products\_detail\_url){

product\_detail\_page<-read\_html(product\_detail\_url)

**#lay ten san pham**

product\_name <- html\_text(html\_node(product\_detail\_page,"#detail-name"))

productName <- c(product\_name)

**#duong dan anh cua san pham**

product\_img <- html\_attr(html\_node(product\_detail\_page,".z-fea img"),"src")

productImage <- c(product\_img)

productImage <-c(str\_c("https://tainghe.com.vn/",productImage))

**#gia ban cua san pham**

product\_final\_price<-html\_text(html\_node(product\_detail\_page,".price\_config"))

**#loai bo don vi 'đ'**

product\_final\_price<-str\_sub(product\_final\_price,0,str\_length(product\_final\_price)-2)

if(is.na(product\_final\_price)){

product\_final\_price<-0

}else{

**#loai bo dau cham**

productPrice <- c(product\_final\_price)

productPrice <- str\_replace\_all(productPrice, "[[:punct:]]" , "")

}

**#lay ma san pham**

productId <- c(str\_c("PRODUCT",i))

#lay duong link cua danh muc

categoryHref <- c(category\_url)

temp <- data.frame(productId, productName, productPrice, productImage, categoryHref)

SanPham <- rbind(SanPham,temp)

i<-i+1

}

}

}

* + 1. Lấy các node và thông tin của sản phẩm theo từng loại

Tạo vòng lặp **for** chạy từng đường dẫn của từng loại sản phẩm. Thêm **‘https://tainghe.com.vn’** vào trước mỗi đường dẫn loại sản phẩm để hệ thống có thể xử lý kết nối.

Lấy các node cần lấy của các sản phẩm để lấy thông tin.

Xử lý thông tin lấy được như loại bỏ các dấu ‘.’, ‘đ’ trong giá sản phẩm, thêm **‘https://tainghe.com.vn’** vào trước giá trị của đường dẫn hình ảnh sản phẩm để có thể truy cập.

* + 1. Tạo thông tin khóa ngoại giữa 2 bảng loại sản phẩm và sản phẩm

**#join bang san pham voi loai san pham de tim ma danh muc**

SanPham.lookup <- data.frame(LoaiSanPham[3],LoaiSanPham[1])

SanPham<-merge(SanPham, SanPham.lookup)

SanPham <- SanPham %>% select(productId, productName, productPrice, productImage, categoryId)

LoaiSanPham <- LoaiSanPham %>% select(categoryId:categoryName)

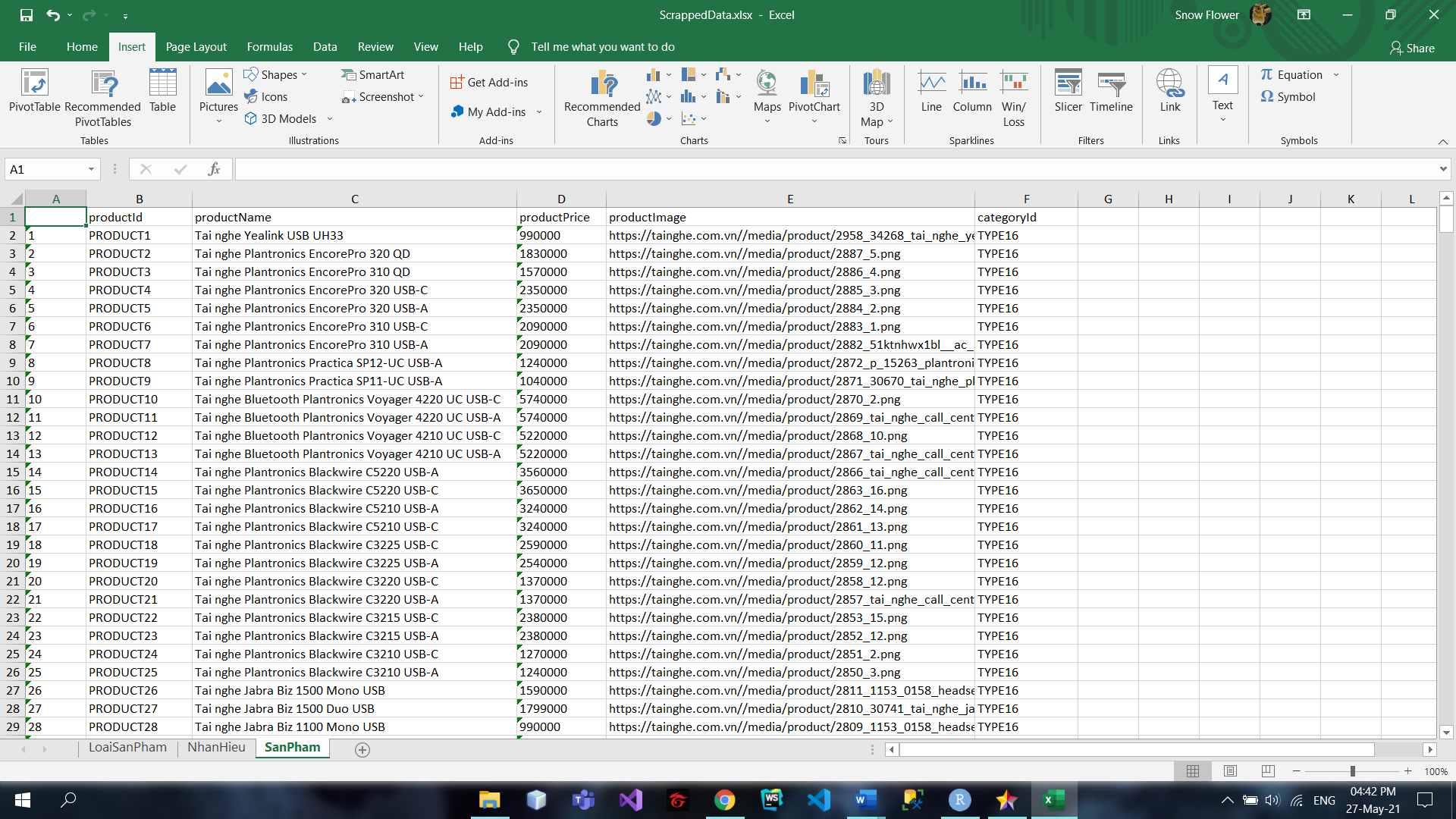
* + 1. Xuất 3 bảng dữ liệu vào 3 sheet khác nhau

write.xlsx(LoaiSanPham,file= "D:\\HOCTAP\\220\\ChuyenDeCSDL\\R\\HeadPhone\\ScrappedData.xlsx",sheetName="LoaiSanPham",append=TRUE)

write.xlsx(NhanHieu,file= "D:\\HOCTAP\\220\\ChuyenDeCSDL\\R\\HeadPhone\\ScrappedData.xlsx",sheetName="NhanHieu",append=TRUE)

write.xlsx(SanPham,file= "D:\\HOCTAP\\220\\ChuyenDeCSDL\\R\\HeadPhone\\ScrappedData.xlsx",sheetName="SanPham",append=TRUE)

* + 1. Kết quả dữ liệu



* Link project: https://github.com/thanhluan17/CDCSDL