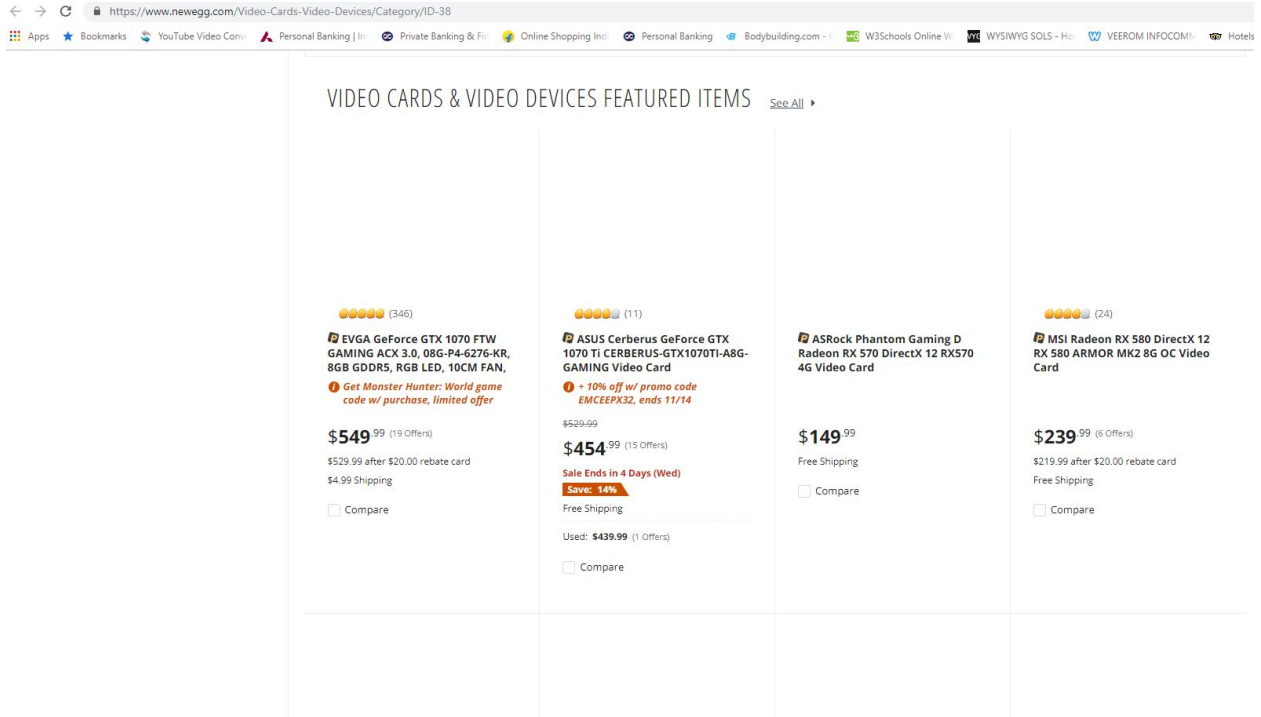


Web Scraping using R

I am illustrating code to extract prices and information on current products shown in website below:



Code:-

```
library(rvest)
library(stringr)
now <- Sys.time()

# url to scrape, then download page
url <- "https://www.newegg.com/Video-Cards-Video-Devices/Category/ID-38"
webpage <- read_html(url)

#####
# web scraping
#####

#####
# feature: card name
#####
card_name <- webpage %>% html_nodes(".item-title") %>% html_text()
```

```
#####
# feature: current price
#####
cur_price <- webpage %>% html_nodes(".price-current strong") %>% html_text()

org_price <- webpage %>% html_nodes(".price-was") %>% html_text(trim=TRUE)

# substring search for price, using regular expression.
needle <- "\\d{1,}\\|\\.\\d{1,}"
indexes <- str_locate(string = org_price, pattern = needle)
indexes <- as.data.frame(indexes)
org_price <- str_sub(string=org_price, start = indexes$start, end = indexes$end)

rate.pid <- webpage %>% html_nodes(".item-rating") %>% html_attr("href")
# format: <url><"Item="><pid><'><stuff>
rate.pid.split <- str_split_fixed(rate.pid, pattern = "Item=", n=2)
# result: [1] [2]
#         <url>  <pid><'><stuff>
rate.pid.split <- str_split_fixed(rate.pid.split[,2], pattern="&", n=2)
# result: [1] [2]
#         <pid>  <stuff>
rate.pid <- rate.pid.split[,1]

# rating
rating <- webpage %>% html_nodes(".item-rating") %>% html_attr("title")
# result: <string><+\\s><rating>
rating <- str_split_fixed(string = rating, pattern="\\+\\s", n = 2)[,2]
# result: [1] [2]
#         <string\\s> <rating>
rating_df <- as.data.frame(cbind(rate.pid, rating))

# combine

graphics_cards <- as.data.frame(card_name)
graphics_cards$scrape_date <- now
graphics_cards$cur_price <- current_price
graphics_cards$org_price <- org_price
graphics_cards$rating <- rating

na.org_price <- is.na(graphics_cards$org_price)
graphics_cards[na.org_price,"org_price"] <- graphics_cards[na.org_price,"cur_price"]

# cast into numeric
graphics_cards$org_price <- as.numeric(graphics_cards$org_price)
graphics_cards$cur_price <- as.numeric(graphics_cards$cur_price)
```

```
# sales price - current price = sales discount
graphics_cards$sales_amt <- graphics_cards$org_price - graphics_cards$cur_price

graphics_cards$discount <- graphics_cards$sales_amt / graphics_cards$org_price

#####
# feature: on_sale
#####
# logic:  if discount price as a percentage of the original price is higher than
#         a certain percentage threshold, mark as being on sale
# key:  0 = not on sale
#       1 = on sale
threshold <- 0.03
graphics_cards$on_sale <- 0
graphics_cards[graphics_cards$discount > threshold, "on_sale"] <- 1
```

Output:=>

Environment		History	Connections
Import Dataset			
Global Environment			
Data			
graphics_cards	12 obs. of 4 variables		
indexes	12 obs. of 2 variables		
rate.pid.split	chr [1:10, 1:2] "N82E16814487259" "N82E1681412623..."		
rating_df	10 obs. of 2 variables		
webpage	List of 2		
Values			
card_name	chr [1:12] "EVGA GeForce GTX 1070 FTW GAMING ACX 3...."		
cur_price	chr [1:12] "549" "454" "149" "239" "499" "209" "249..."		
na.org_price	logi [1:12] TRUE FALSE TRUE TRUE TRUE TRUE ...		
needle	"\\d{1,}\\..\\d{1,}"		
now	2018-11-11 19:49:12		
org_price	chr [1:12] NA "529.99" NA NA NA NA NA NA NA NA NA NA		
rate.pid	chr [1:10] "N82E16814487259" "N82E16814126231" "N82..."		
rating	chr [1:10] "5" "4" "4" "5" "4" "4" "4" "5" "5" "5"		
threshold	0.03		
url	"https://www.newegg.com/Video-Cards-Video-Devices/C..."		