TBA2105 Web Mining
2022/23 Sem 1
Individual Assignment 2

Submission Information

- This assignment contributes 15% to the final course grade. The total mark for this assignment is 15.
- Deadline: **6 Nov 2022 2359hrs (Sunday 11:59pm) (workbin will autoclose)**
- Upload a single zip file to LumiNUS workbin (Deliverables Submission > Assignment 2)
    - The single zip file should contain:
        - One or more R scripts for scraping and analyzing the data
        - Report (in docx or pdf format)
        - Dataset (in CSV format) if not large
- **You should refrain from discussing the assignment with anyone and should only submit original work. You will be subjected to disciplinary actions if you attempt to copy or assist in copying.**
- **Please ensure that you have written your name and matric number in the document**

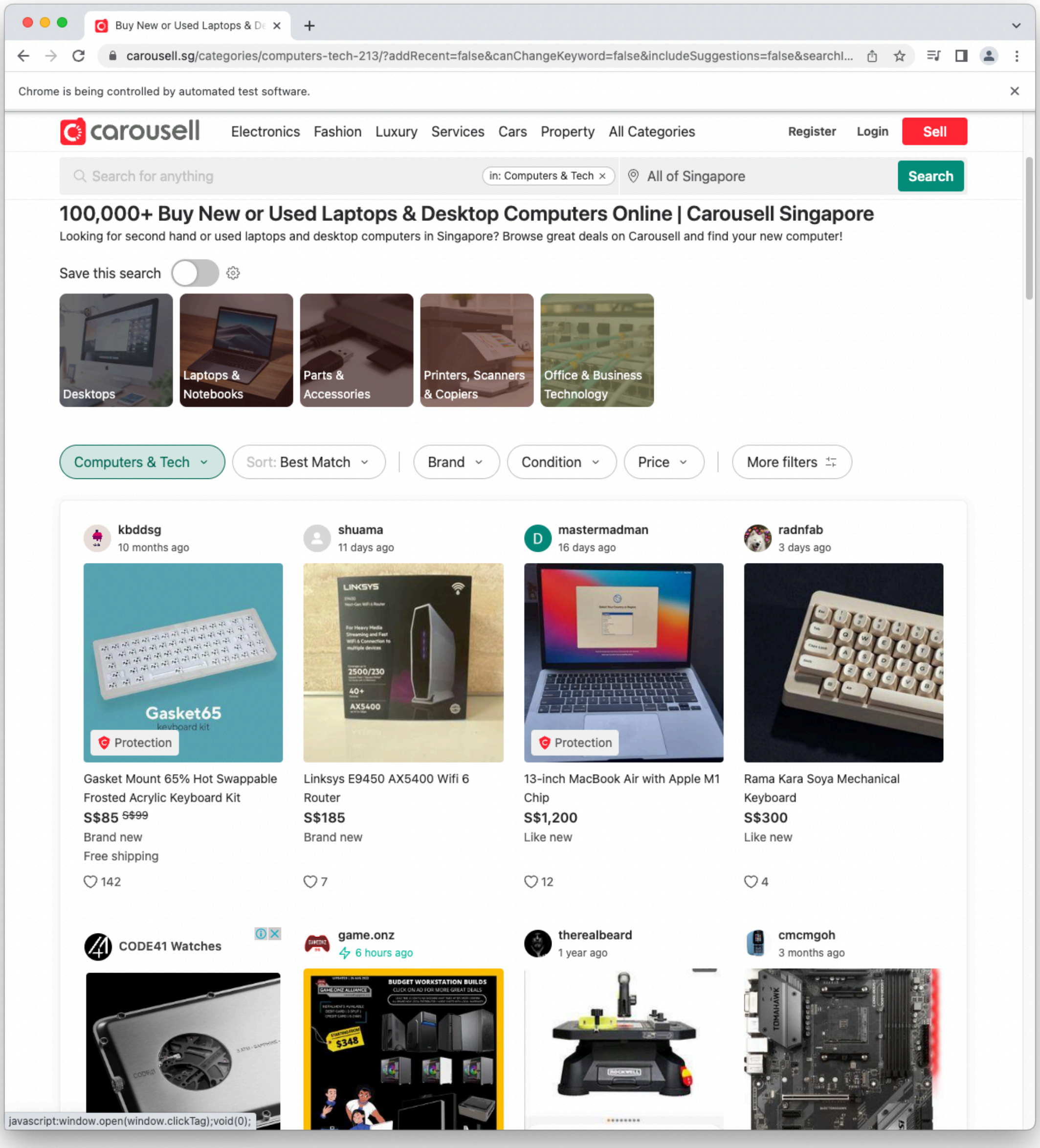
Learning Objectives

- Use R to perform web scraping on a real website
- Analyze data in a critical manner and come up with useful findings

1. This assignment is meant to assess how you apply what you have learned so far in the course. You will be performing web scraping on a well-known local mobile app focusing mainly on selling used products (https://www.carousell.sg/). Note that while most users access the services through the Carousell mobile app, the listing information is also accessible through the Carousell website which you will be using for doing the web scraping.

2. You will be mining data from the Carousell website and generating useful insights derived from this data. The exact details of what to scrape and analyze from the website are up to you. You can refer to the following for some inspiration (this list is not exhaustive and you are also not required/expected to choose from the following). You are also not restricted (and encouraged) to derive as many insights as possible.
    a. What is the distribution of the prices of products in a category?
    b. What is the nature of product listing in a category vs another category?
    c. Is there any relationship between the number of likes vs the price?
    d. Is there any relationship between the number of likes vs the nature of the item?
    e. etc

3. Just an illustration, suppose you are generating insights of a category e.g. https://www.carousell.sg/categories/computers-tech-213/, we start by mining all (or substantial amount e.g. 500 listings) of the products entries in the category. As a starting point, you might want to mine the seller, title, price, and number of likes from each product listing. As the website loads data dynamically, you will most likely need to use RSelenium to do the web scraping. A snapshot of the dataset might look like the screenshot below. You do not need to follow the same columns and format given below and are encouraged to consider including other interesting information.



Buy New or Used Laptops & De ×

carousell.sg/categories/computers-tech-213/?addRecent=false&canChangeKeyword=false&includeSuggestions=false&searchl...

Chrome is being controlled by automated test software.

100,000+ Buy New or Used Laptops & Desktop Computers Online | Carousell Singapore

Looking for second hand or used laptops and desktop computers in Singapore? Browse great deals on Carousell and find your new computer!

Save this search

kbddsg
10 months ago
Protection
Gasket Mount 65% Hot Swappable Frosted Acrylic Keyboard Kit
S$85 S$99
Brand new
Free shipping
142

shuama
11 days ago
Linksys E9450 AX5400 Wifi 6 Router
S$185
Brand new
7

mastermadman
16 days ago
Protection
13-inch MacBook Air with Apple M1 Chip
S$1,200
Like new
12

radnfab
3 days ago
Rama Kara Soya Mechanical Keyboard
S$300
Like new
4

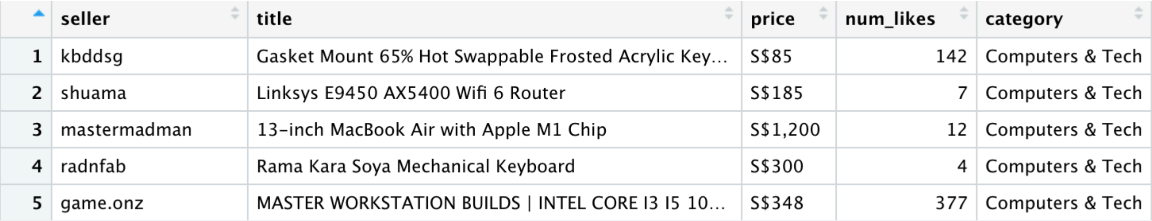
CODE41 Watches

game.onz
6 hours ago

therealbeard
1 year ago

cmcmgoh
3 months ago

javascript:window.open(window.clickTag);void(0);



	seller	title	price	num_likes	category
1	kbddsg	Gasket Mount 65% Hot Swappable Frosted Acrylic Key...	S$85	142	Computers & Tech
2	shuama	Linksys E9450 AX5400 Wifi 6 Router	S$185	7	Computers & Tech
3	mastermadman	13-inch MacBook Air with Apple M1 Chip	S$1,200	12	Computers & Tech
4	radnfab	Rama Kara Soya Mechanical Keyboard	S$300	4	Computers & Tech
5	game.onz	MASTER WORKSTATION BUILDS | INTEL CORE I3 I5 10...	S$348	377	Computers & Tech

4. Please include the dataset(s) (in CSV) format that is mined. If the dataset is too large, you can just take a sample of the dataset instead.

5. Using the dataset(s) that you obtained, write a report. Some of the information (not limited to) can include:
    a. Highlight the problem statement/overall analysis you are focusing on for this study.

b. Describe the design and structure of the web scraper **written in R**.
c. Discussion of the challenges encountered during the web scraping process and strategies adopted to tackle these challenges.
d. Where applicable, provide a set of charts (e.g. ggplot2 charts, wordcloud, etc), statistical measures (e.g. mean, median, etc), or models (e.g. regression models) that are generated.
e. Discussion of the results from d and explain how this information (where applicable) might be helpful for Carousell, a buyer of Carousell, a seller of Carousell, or a competitor to Carousell.

6. Please take note that this assignment is mainly for educational purposes. You will be liable for the way how you use the data according to Carousell's terms of use. We will not be liable for how you use the data.

Grading Criteria

Weightage 15% of the course grade

- 2% report organization and presentation
- 5% analysis and discussion
- 8% codes and showing relevant results.