TBA2105 Assignment 2

Name: NOEL LIM XIAN
Student No: E0493357
Date: 8 Nov 2022

References

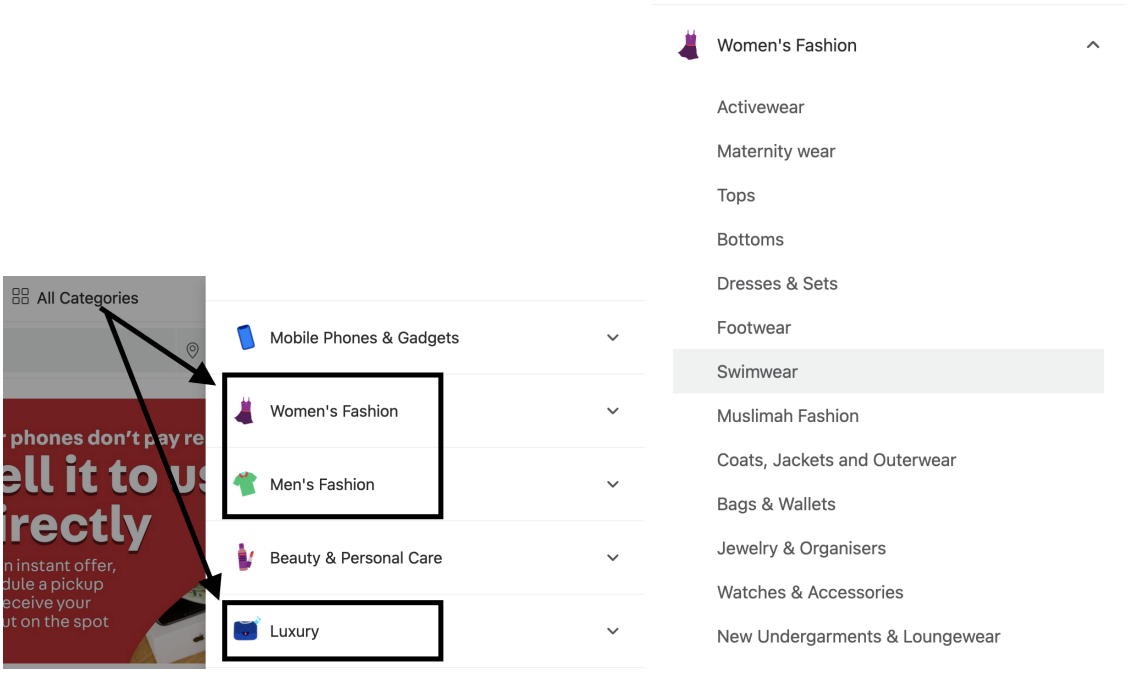
assignment_2_extraction.r - Data scraping procedure from https://www.carousell.sg/
dataset.csv - Yield from scraping website
assignment_2_analysis.r - Data analysis performed on *dataset.csv*
dataset.csv - Summary table
report.pdf - This report.

Introduction

Carousell is a platform for customer-to-customer and business-to-customer listings. It has a strong presence in the Asia-Pacific Region. Singapore's local website of Carousell is at https://www.carousell.sg/ and the data collected from it will be analyzed in this report.

Our focus will be on fashion items, which are indicated under 3 main categories - Women's Fashion, Men's Fashion and Luxury. Main categories contain subcategories which have their own page listing.

Categories (left) and Subcategories (right)



R and its packages are used to interact with the website and also perform data wrangling and analysis.

Extraction Steps

The extraction steps are written to be automated such that we can perform web crawling from a fresh environment on every run. This includes restarting the browser, and only access unauthenticated routes without website personalization (cookies etc.) to avoid user bias. Our objective is to obtain the url locations to the pages of subcategories, then sequentially visit these pages to view item listings and retrieve the relevant details.

Initialization

1) Terminate existing browser processes and open a new browser via RSelenium driver.
2) Go to "https://www.carousell.sg" and coerce language to English.

Url Fetching

3) Expand “All Categories” modal to view the list of categories.
4) Expand the fashion categories to view subcategories and obtain their page redirect link.

5) For each subcategory page, primary sorting is set to recency and we obtain the price, condition and name of the first 100 selling items.
   * *In order to step through and display more items, the browser is scrolled to the bottom to ensure the “Get More Results” button is rendered. The button is then clicked until at least 100 items are displayed.*

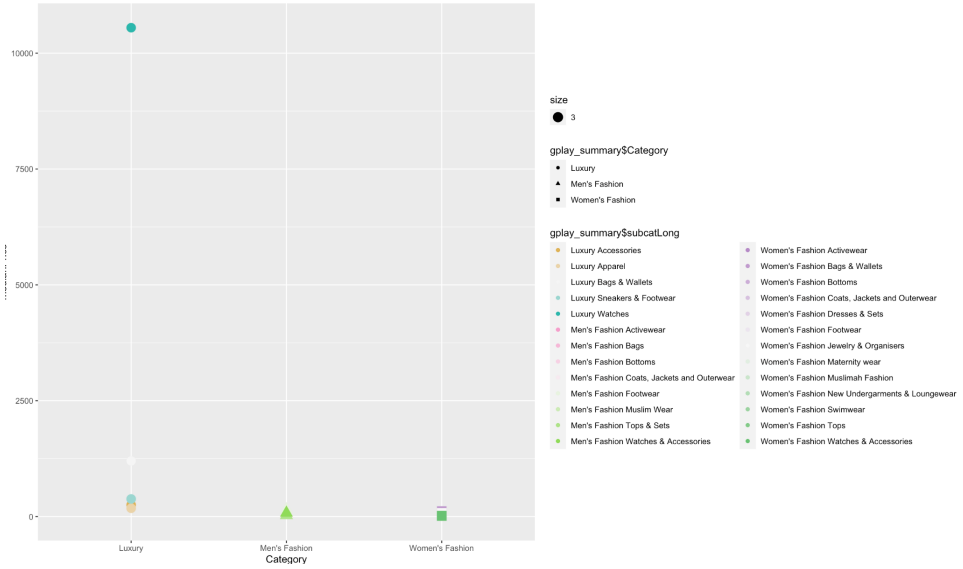
Processing Steps

6) The price, which is denoted in SGD (i.e “S$100”), is normalized to numerical value.

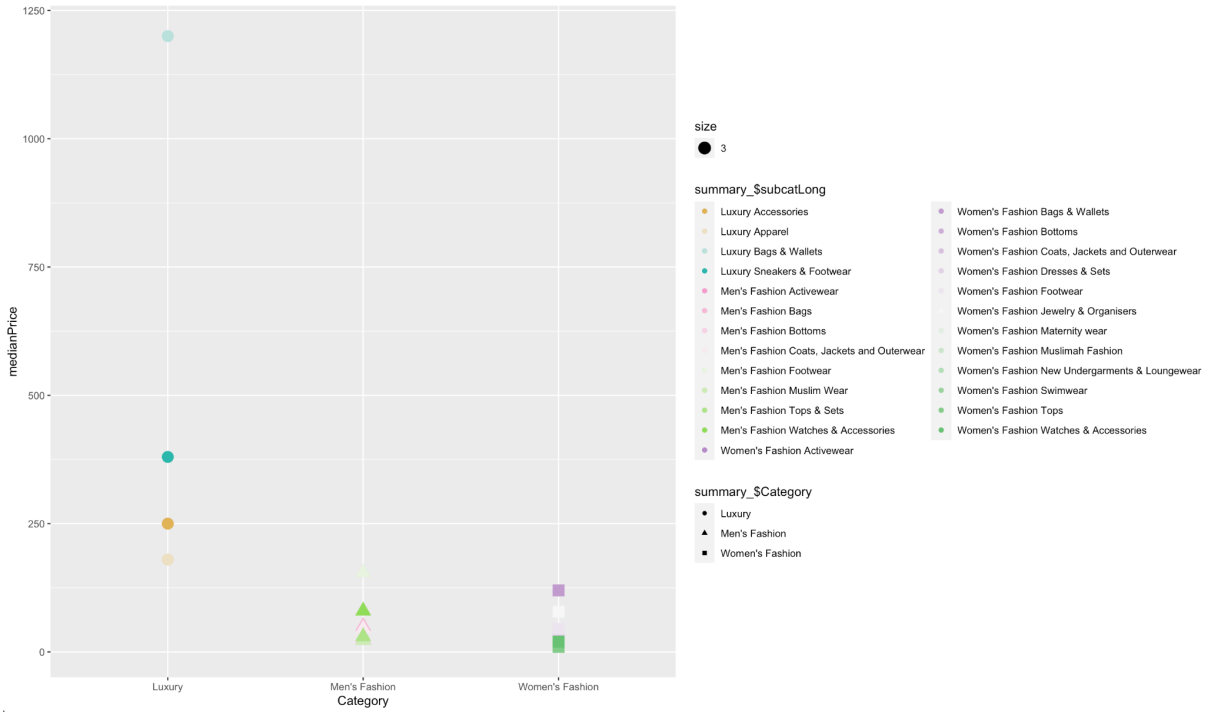
Analysis

Median Price of Subcategories

We aggregate the median of each subcategory in a new table. Except for “Watches” and “Bag & Wallets” of category Luxury, the median prices of the listing are within $0 to $2500.



Eliminating the above-mentioned outliers, it is observed that minimum median prices in Luxury subcategories is more than both Women's and Men's Fashion subcategories.



END