Li Tao

850001413

GDDA612\_Assessment 1\_Project

Table of Contents

[PART A - Data Collection 2](#_Toc170483525)

[Task 1: Identify Data Sources 2](#_Toc170483526)

[Task 2: Web Scraping 4](#_Toc170483527)

[a) Web scraping for 3 data sources and storing in datasets 4](#_Toc170483528)

[b) Ethical standards and data privacy compliance in web scraping 12](#_Toc170483529)

[c) Data collection and managing cultural perspectives of Tikanga, Whakapapa, and Pepeha. 13](#_Toc170483530)

[PART B - Data Preparation and Cleansing 14](#_Toc170483531)

[Task 3: Data Preparation and Cleansing 14](#_Toc170483532)

[Task 4: Documentation 22](#_Toc170483533)

[a) Data preparation and cleaning for enhanced marketing strategies 22](#_Toc170483534)

[b) 5 Challenges faced in data preparation and cleaning and their solutions 24](#_Toc170483535)

[PART C - Data Importation 26](#_Toc170483536)

[Task 5: Store Datasets 26](#_Toc170483537)

[Task 6: Merge Data 27](#_Toc170483538)

[Task 7: Indexing 29](#_Toc170483539)

[a. Set a specific column as the index. 29](#_Toc170483540)

[b. Reset the index to the default integer-based index. 29](#_Toc170483541)

[c. Create a new data frame by selecting rows based on a conditional index. 30](#_Toc170483542)

[d. Perform multi-level indexing by setting multiple columns as the index. 30](#_Toc170483543)

[Task 8: Sorting 31](#_Toc170483544)

[Task 9: Summary Statistics 32](#_Toc170483545)

[Task 10: Slicing 33](#_Toc170483546)

[Task 11: Data Import 34](#_Toc170483547)

# PART A - Data Collection

## Task 1: Identify Data Sources

Identify and describe three different online data sources

­

In this paper, I focused on the retail and cosmetics field. Implementing the chosen scenario into practice, I selected the skincare sections of three well-known skincare and beauty websites: Sephora, Lookfantastic, and Mecca. The plan was to analyse the skincare products using web scraping. This helped me understand their market positioning, pricing strategy, sales tactics, and customer reviews. Through inspecting the webpage, I found some discrepancies on the website regarding layout and formatting. If there were any null data or duplicated data after the web scraping, I would make necessary adjustments and clarifications in the other tasks.

The web pages where I performed web scraping are shown below:

1. Sephora skincare web page:

Sephora is a global retailer specializing in beauty products. The website offers a wide range of skincare products, from luxury to affordable options, providing detailed product descriptions, and customer reviews, catering to different skincare concerns and preferences, and featuring frequent promotions.

A screenshot of a website

Description automatically generated

1. Lookfantastic web page:

Lookfantastic is the top beauty retailer in Europe, offering a diverse option of skincare products from both well-established and niche brands. The website often runs exclusive collaborations and limited-edition collections.

A screenshot of a website

Description automatically generated

1. Mecca web page:

Mecca is an Australian beauty retailer renowned for its curated selection of products. The website features a blend of luxury and unique brands, as well as Mecca-exclusive gift sets. Mecca offers high-quality skincare products alongside personalized beauty advice.

A screenshot of a product

Description automatically generated

## Task 2: Web Scraping

### a) Web scraping for 3 data sources and storing in datasets

I used Python programming provided by Google Colab to perform web scraping for this task. On the first website, Sephora, I tried several methods to locate the review counts and ratings, I showed the screenshots below. On the other web pages, I looped and stored the scraped data directly. I clarified the steps with screenshots.

#### Screenshot of applying web scraping on the Sephora web page:

1. Import relevant libraries and use the ‘requests’ to get a response from the specified ‘url’.

Use the ‘BesutifulSoup’ to parse the HTML content from the website into ‘soup’ object.

A screenshot of a computer

Description automatically generated

1. Locate the target element and store it in the ‘table’.

Extract the product information by finding all elements within the ‘table’.

Got a list of tag objects representing each product.

A screenshot of a computer code

Description automatically generated

1. The first approach is to extract the rating result.

Find the rating information and store it in ‘rating’.

Retrieve the ‘style’ attribute.

Extract the percentage value embedded in the ‘style’ attribute.

Convert the value to float to handle decimals and then convert to integer for easier handling.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

1. The second approach is to extract the review counts.

I used the same technique in this part.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Loop to store all the scraped data

With an understanding of the previous steps, I rewrote the Python script to initialize an empty list named ‘skincare’ outside the loop to store all the scraped data in categories accordingly. The loop iterated through each product extracting specific details. The details were appended to the ‘skincare’ list.

A screenshot of a computer program

Description automatically generated

Show the result:

A screenshot of a computer

Description automatically generated

1. Import the extracted data into the data frame

A screenshot of a computer

Description automatically generated

#### Screenshot of applying web scraping on the Lookfantactic web page:

I performed the same steps as on the Sephora web page.

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

Import the extracted data into the data frame.

A screenshot of a product list

Description automatically generated

#### Screenshot of applying web scraping on Mecca web page:

Again, I performed the same steps as on the previous web page.

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screen shot of a computer program

Description automatically generated

A screenshot of a phone

Description automatically generated

A screenshot of a computer

Description automatically generated

### b) Ethical standards and data privacy compliance in web scraping

When selecting a website, my first step was to carefully read its privacy policy to ensure transparency and publicity in the data collected.

During the web scraping, I only gathered relevant information essential to my paper, retail and cosmetics research, such as product names, functions, prices, ratings, and review counts. I avoided any sensitive details like guest identities on the website.

In addition, all the data captured would be maintained strictly confidential and secure, ensuring no unauthorized disclosure occurs.

At last, I believe it’s crucial to control the frequency of website scraping. This approach not only lessens the load on the website and mitigates any negative impact on its performance but also safeguards my activity from being mistaken for malicious behaviour.

### c) Data collection and managing cultural perspectives of Tikanga, Whakapapa, and Pepeha.

Life in New Zealand is intricately connected to Māori culture. Respect for and protection of Māori society permeates all aspects of New Zealand life. The data collection and processing will naturally be carried out with utmost respect and consideration for these Māori cultural values.

“Tikanga” refers to the rules, traditions, and customs of Māori society and culture, covering a variety of values and norms. When conducting web scraping, the entire process must respect individuals’ rights and autonomy. In situations involving the disclosure of personal information, individuals must be informed of their right to know, and the purpose of data collection must be clearly explained.

“Whakapapa” refers to the history and inheritance of the family, and embodies the concepts of connection, synthesis, and wholeness. When it comes to data analysis, the source of the data, the method of collection, and the alignment with the purpose need to be reasonable and transparent. This ensures the facilitation of the analysis and understanding.

“Pepeha” refers to a personal relationship with family, land, and community, which can be understood as a cultural and spiritual connection. In data analysis, understanding the “Pepeha” of data involves respecting an individual's identity and cultural affiliation, avoiding sensitive information, and ensuring the accuracy and reliability of the data.

# PART B - Data Preparation and Cleansing

## Task 3: Data Preparation and Cleansing

Implement data cleansing methods using scripts to prepare all three datasets for further analysis.

#### Sephora data frame:

1. Check the null value and the data type

0 null value in the dataset. All the variables are in ‘object’ type.

A screenshot of a computer program

Description automatically generated

1. In the ‘price’ column, replace the $ sign with space, and convert its type to numeric.

In the ‘rating’ and ‘review\_counts’ columns, extract the numeric values. Replace the empty value with 0. It is acceptable that there are no comments or ratings on some products.

A screenshot of a computer

Description automatically generated

1. Check the data type and store the dataset in a CSV file for further analysis

A close-up of a computer screen

Description automatically generated

1. Check the outlier in the numeric columns.

I decided to retain the outliers because these values in ‘price’, ‘rating’, and ‘review\_counts’ fall within a genuine and acceptable range. Products can vary widely in exposure and popularity, with some garnering significant attention while others may remain unnoticed in the market.

A screenshot of a computer code

Description automatically generated

Show the result:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

#### Lookfantastic data frame:

1. Using regular expressions to extract the ‘Stars’ and ‘Reviews’ numbers.

Split the ‘review’ column, and assign the values to new columns ‘star’ and ‘review\_counts’ in the data frame.

A screenshot of a product list

Description automatically generated

A screenshot of a computer program

Description automatically generated

1. Define the rates and exchange the pounds for NZ dollars.

A screenshot of a computer

Description automatically generated

1. Scale the ‘stars’ column.

A screenshot of a computer program

Description automatically generated

1. Split the ‘product\_name’ column by indexing and extracting the last four words.

Store the extracted values in new columns ‘function’ and ‘name’.

A screenshot of a computer

Description automatically generated

1. Drop the column ‘ product\_name’ because it is a duplicated variable

A screenshot of a computer

Description automatically generated

1. Rename and reorder the columns

Save the data frame in a CSV file for further analysis

A screenshot of a computer

Description automatically generated

#### Mecca data frame:

1. Extract the price

Extract and return the last price in the ‘price’ column because some products’ price range started from the sample price which is not the normal retail price.

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

1. Drop the $ sign and change the data type to float for ‘price’ column

A screenshot of a computer

Description automatically generated

1. Replace N/A values with 0 in the ‘rating’ and ‘review\_counts’ columns and convert the data type to integers

A screenshot of a computer code

Description automatically generated

1. Check and drop duplicated columns

A screenshot of a computer program

Description automatically generated

1. Save the data frame in a CSV file for further analysis



## Task 4: Documentation

### a) Data preparation and cleaning for enhanced marketing strategies

* Sephora skincare dataset:

1. Check null and found null is 0.
2. Check the data type, all are object data types.
3. Remove the $ sign in the ‘price’ column and convert the column to the numeric type.
4. Extract the numeric values from the ‘rating’ column: select the ‘rating’ column, apply the lambda function to each element ‘x’ in the ‘rating’ column, and use the if isinstance function to check if ‘x’ is of type ‘list’, convert ‘x’ to integer, otherwise apply 0 to ‘x’ as the value.
5. Extract numeric values from the ‘review\_counts’ column: apply the same technique plus one step to remove the parentheses before converting to the integer.

* Mecca skincare dataset:

1. Check the dataset null value and data type.
2. Extract the last price in the price range as the lower price may refer to a sample product which is not an accurate price.
3. Drop the $ sign in the ‘price’ column and convert the column to float data type.
4. Convert the ‘rating’ and ‘review\_counts’ columns to integer data type. Replace the N/A value with 0, as it means no comments yet.

* Lookfantastic skincare dataset:

1. Split the ‘review’ column into ‘stars’ and ‘review\_counts’ columns. Drop the ‘review’ column.
2. Define the rates and exchange the price to NZ dollar currency. The website uses the British pound as the price currency, so the currency exchange is applied here.
3. Scale the ‘stars’ column. On other websites, they utilize a system based on 100 points, whereas, on the Lookfantastic website, a 5-star system is employed. I converted the 5-star ratings into a 100-point scale to ensure consistency for subsequent operations within the same measurement framework.

After completing the data preparation and cleaning process, all three datasets now share the same format and scale structure. Each dataset includes feature columns such as ‘product name’, ‘product function’, ‘price’, ‘rating’, and ‘review counts’. The columns ‘price’, ‘rating’, and ‘review counts’ are represented in numeric data type, while the remaining columns are represented in object data type. Each row within each dataset represents the information for one product.

According to the dataset, several marketing strategies could be effectively applied:

* Function: Positioning the right product for specific target audiences is crucial. For example, moisturizing products can be tailored to different age and gender groups. The sunscreen can be strategically promoted ahead of summer, highlighting its SPF and UV protection levels.
* Price: Setting up the competitive price involves analysing competitors’ price ranges and aligning them with the website’s branding.
* Rating and Review\_counts: These columns can be used as marketing highlights. Products with high ratings and numerous reviews enhance their reputation and increase consumer confidence in purchasing decisions. Positive feedback can be prominently featured on product pages and social media. Based on the data, promotions and incentive programs can be designed to drive sales. Understanding the sales and customer feedback enables the prediction of market trends. These insights aid in the development of precise marketing strategies and effective inventory management plans.

### b) 5 Challenges faced in data preparation and cleaning and their solutions

1. Extract the rating number without disturbance

On the Sephora website, I wanted to extract the rating results, and it was shown in percentage. I initially encountered challenges where the extracted data included HTML tags and additional metadata such as the ‘maxStars’ value in. To overcome this, I found a method to precisely isolate the actual rating number. I split what was shown after the highlighted percentage and split what was shown before the ‘%’ sign. After that, I successfully obtained the precise rating number.

2. Extract the plain text for subsequent processing

When using Beautiful Soup to scrape website content, the data is typically returned in HTML format. This means when I assessed the variables and printed them, they included tags, attributes, and text within those tags. To obtain clean text without HTML tags from web elements, I used the ‘.text’ method. This allowed me to extract only the plain text content, stripping away any HTML tags and making it easier and more suitable for subsequent processing and analysis.

3. Scale the rating system to ensure consistency across three datasets

On the Lookfantastic website, product ratings are based on a 5-star system, whereas other websites use a 100-point scale. To ensure consistency across different measurement frameworks, I converted the 5-star ratings into a 100-point scale. During this process, I encountered two interpretations regarding how to map stars to points: one considers 0 points as 1 star, while the other considers 0 points as 0 stars. To clarify this, I conducted further research on the Lookfantastic website by examining customer reviews. I discovered that the lowest rating available is 1 star, indicating that 20 points correspond to 1 star. Meanwhile, on other websites using the 100-point scale, the lowest rating also corresponds to 1 star, suggesting that each star represents 20 points in this conversion. This approach ensures that ratings from different websites are aligned on a common scale, facilitating consistent analysis and comparison across products.

4. Monetary exchange

On the Lookfantastic website, product prices are quoted in British pounds (£). To compare and analyze prices across three websites using a consistent monetary standard, I needed to establish an exchange rate from pounds to New Zealand dollars (NZD). I conducted online research to find the current exchange rates and used mathematical calculations to apply the conversion accurately. By defining a reliable exchange rate, I ensured that prices from different websites could be compared on an equal footing in terms of their monetary value in NZD. This approach enabled a fair and meaningful analysis of pricing strategies and trends across the platforms.

5. Extract numbers from brackets with replacing N/A values

In the Sephora dataset, I needed to extract the numbers from the ‘rating’ and ‘review counts’ columns, which were formatted as objects containing brackets ‘[]’ and parentheses ‘()’. Initially, I attempted to use the ‘strip()’ method to remove spaces and characters, but the brackets and parentheses persisted. Directly converting these values to integers also failed because some cells might contain ‘N/A’ values, indicating no customer ratings. To address this, I applied a lambda function to each element ‘x’ in the ‘rating’ column. Within the lambda function, I used the ‘isinstance’ method to check if ‘x’ was of type ‘list’. If so, I converted ‘x’ to an integer; otherwise, I assigned ‘0’ to ‘x’ to handle cases where ratings were absent due to no customer reviews. This approach ensured that I avoided errors caused by ‘N/A’ values and correctly interpreted products without ratings on the website. After successfully implementing this for the ‘rating’ column, I applied the same technique to the ‘review\_counts’ column with an additional step: before converting the values, I removed parentheses to prevent errors during the conversion process.

# PART C - Data Importation

## Task 5: Store Datasets

Python script to save cleaned datasets in CSV format

I used the Pandas library in Python to save the three DataFrames into CSV files.

I defined the CSV file path in my Google Drive under the ‘612 course/612 assessment 1’ folder.

Use ‘df.to\_csv’ to save the DataFrame to a CSV file without including the index column.

Please see the screenshots below:

Sephora:

A close-up of a computer screen

Description automatically generated

Lookfantastic:

A screenshot of a computer

Description automatically generated

Mecca:



## Task 6: Merge Data

Python script to extract and specify merge operations for three datasets

In this task, I created a new notebook named ‘612 A1 PartC’ and imported relevant libraries.

1. Read and import the three datasets into separate DataFrames, named ‘df1’, ‘df2’, and ‘df3’.

A screenshot of a computer

Description automatically generated

1. Browsing the data frame by showing a few rows in each dataset

A screenshot of a computer

Description automatically generated

1. Merge the data frames by using the ‘concat’ function

To combine these three datasets without losing data or causing null values, it’s appropriate to use the concatenation method since these features share the same column names, categories, and data types. I concatenated the DataFrames along rows, stacking one DataFrame below the other.

After the data merged, I got a new DataFrame named ‘merged\_df’ with 213 rows and 5 columns.

A screenshot of a computer

Description automatically generated

## Task 7: Indexing

Python script to perform indexing on a merged DataFrame

I used the merged data frame to perform the below tasks.

A black text on a white background

Description automatically generated

### Set a specific column as the index.

I added a new column named ‘product\_id’ and set it as the index.

The ‘inplace = True’ argument modifies the DataFrame in place, meaning it directly affects ‘df’ without creating a new DataFrame.

A screenshot of a computer program

Description automatically generated

### Reset the index to the default integer-based index.

The ‘drop = True’ ensures the old index is dropped rather than being added as a new column. This resets the index to default integer-based indices starting from 0.

A screenshot of a computer

Description automatically generated

### Create a new data frame by selecting rows based on a conditional index.

I created a new data frame by selecting the rows where the ‘price’ column exceeds 50.

Out of 213 products, 112 products have prices over 50.

A screenshot of a computer

Description automatically generated

### Perform multi-level indexing by setting multiple columns as the index.

I selected the ‘product\_name’ and ‘function’ columns as the index.

After this step, I can locate specific products by finding their name and function.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Task 8: Sorting

Sort the merged Data Frame based on the values in ascending and descending order in a specific column.

I sorted the ‘price’ column in ascending order to find the products with the lowest price.

A screenshot of a computer

Description automatically generated

I sorted the ‘rating’ column in descending order to find the products with the highest rating.

A screenshot of a computer

Description automatically generated

## Task 9: Summary Statistics

Perform summary statistics and identify key findings

I used the Pandas built-in function ‘describe’ to perform and display the summary statistics in the data frame.

The data frame has three numeric columns: ‘price’, ‘rating’, and ‘review\_counts’, each containing 213 values, meaning each column has 213 data points.

A screenshot of a computer

Description automatically generated

* Price:

The average price of the items is 60. The price ranges from a minimum of 2 to a maximum of 461. The standard deviation in price varies significantly around the mean. 50% of the products have prices up to 52. Up to 75% of the products have prices up to 70, indicating most of the products are below 70 while some high-end products increase the mean.

* Rating:

The average rating is 85. The standard deviation in the rating is small, indicating there is no big difference between the rating. Rating ranges from 0 to 100. 75% of the products have a rating of at least 94, showing ratings are generally high.

* Review\_counts:

The average number of reviews per product is 852. Review\_counts vary considerably with a standard deviation of 1722, ranging from 0 to 14867. According to percentiles, 50% of the products have up to 273 reviews, and 75% of the products have up to 802 reviews. The big difference between mean and median indicates the significant levels of popularity among products.

## Task 10: Slicing

Slice and display a specific portion of the data frame with an explanation

I set the two columns ‘product\_name’ and ‘function’ as the index in the previous task. Then I sorted the data frame in ascending order based on the ‘price’ column. In this task, I selected the rows from index 0 to index 98. I selected columns from index 0 to index 1, meaning the first two columns ‘price’ and ‘rating’ were selected, due to ‘product\_name’ and ‘function’ columns indexing.

The slicing operation is useful for focusing on a specific portion of data. In this case, I can quickly find the products with the lower price range and their rating level.

A screenshot of a computer

Description automatically generated

## Task 11: Data Import

Connect Python server to import DataFrame into MongoDB

1. Install ‘pymongo’ package and import relevant libraries

A screenshot of a computer program

Description automatically generated

1. Conver the merged\_df DataFrame to a list of dictionaries

Convert each row into a dictionary where keys are column names and values are row values.

A screenshot of a computer

Description automatically generated

1. Connect to MongoDB and insert data into MongoDB collection ‘612\_A1’

Establish a connection to the MongoDB database.

Select the database named ‘612\_Assessment\_1’ and the collection named ‘612\_A1’.

Insert data into MongoDB and close the connection.

A screenshot of a computer

Description automatically generated

1. Check the data inserted in MongoDB under database ‘612\_Assessment\_1’ and collection ‘612\_A1’

Open MongoDB Compass and refresh the databases. I found the data was successfully inserted into the collection.

A screenshot of a computer

Description automatically generated