HPDP_Project Report.pdf

by Tan Jun Yuan

Submission date: 13-May-2025 09:28PM (UTC-0700)
Submission ID: 2675452190
File name: HPDP_Project_Report.pdf (5.08M)
Word count: 4867
Character count: 30981

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Department of Computer Science
Faculty of Computing

Optimizing High-Performance Data Processing for Large-Scale Web Crawlers

Programme	:	Bachelor of Computer Science (*Data Engineering*)
Subject Code	:	SECP3133
Subject Name	:	High Performance Data Processing
Session-Sem	:	2024/2025-2

Prepared by	:	BERNICE LIM JING XUAN	(A22EC0038)
		KEK JESSLYN	(A22EC0057)
		TAN JUN YUAN	(A22EC0107)
		NAVACHANDER NAVASANTAR	(A22EC0226)
Section	:	01	
Lecturer	:	Dr Mohd Shahizan bin Othman	
Date	:	27-04-2025	

Table of Contents

1.0 Introduction

1.1 Background of the project

In the era of big data, high-performance computing (HPC) plays a critical role in enabling the efficient processing of vast volumes of information from web sources. Web data extraction, or web scraping, has become a fundamental technique for data collection in fields such as e-commerce analysis, sentiment analysis and market research. However, handling large-scale web data introduces significant challenges, including performance bottlenecks, ethical scraping practices and managing crawl delays. To address these challenges, modern scraping systems increasingly incorporate multithreading, multiprocessing and distributed processing techniques to enhance scalability and efficiency.

This project is designed to provide students with practical, hands-on experience in large-scale web data processing using HPC principles. By designing, developing and optimising a web crawler capable of extracting at least 100,000 structured records, students gain insight into real-world technical and ethical challenges associated with web scraping. Furthermore, the project emphasises the importance of system optimisation, particularly through the comparison of different data processing frameworks, thus strengthening critical thinking skills essential for data science professionals.

1.1.1 Web Scraping

This project focuses on collecting and preparing product data from Lazada Malaysia, specifically targeting women-related categories such as Beauty & Skincare, Health & Wellness, Home & Living, Home Appliances, Mother & Baby, Stationery, and Women's Fashion. The main objective is to obtain a clean and structured dataset that can later be used for further analysis or machine learning tasks.

The first step involves web scraping, where product data is automatically collected from the selected subcategories on Lazada. Each subcategory's data is then stored separately in seven Excel files for better organisation. There are a total of 115090 rows of data that have been collected. Once the data is collected, it is uploaded to Google Colab for preprocessing.

1.1.2 Data Processing

In the preprocessing phase, the first task is data integration, where all seven Excel files are combined into a single dataset. To ensure consistency, all string-based fields such as product names are standardised to uppercase formatting. This helps avoid issues caused by inconsistent capitalization during analysis, such as "lotion" and "Lotion" being treated as different items.

Next, we convert important numerical fields like quantity sold and total reviews into numeric data types. This step is crucial because numeric values are required for proper data analysis, such as outlier detection and calculation needed for grouping items into categories.

After ensuring that all data is in the correct format and structure, we handle missing values. For string fields, missing values are filled with "N/A" to clearly indicate unavailable information, while missing numeric fields are filled with 0. This approach ensures that the dataset remains complete without causing errors in future computations. For example, a missing review count is more safely treated as zero than left blank.

Duplicate records are then detected and removed to avoid repetition and ensure data accuracy. Once all the cleaning steps are completed, the final, clean dataset is exported into a CSV file, ready for the optimisation process.

1.1.3 Optimisation Process

The second phase of the project focuses on optimising the cleaned dataset obtained from the initial preprocessing stage. The objective of this phase is to group and analyse products based on pricing tiers, popularity levels, and market performance by location in order to derive meaningful insights and support further analytical tasks.

The first optimisation step involves the categorisation of products into four pricing tiers, which are budget-friendly, affordable, mid-range and premium. Prior to grouping, all records with a price value of RM0 were removed, as such entries are considered illogical or erroneous. Outliers within the price field were then identified. Upon evaluation, these outliers were deemed plausible and were therefore retained. The minimum and maximum prices (excluding outliers) were calculated and used to define the thresholds for each pricing group. Subsequently, all products, including those with outlier prices, were assigned to the appropriate pricing category based on the established range.

The second optimization focuses on product popularity, measured by the total number of reviews. In this stage, outliers were detected and removed to minimise distortion in the analysis. The adjusted minimum and maximum review counts were then used to determine suitable group boundaries. Products were classified into four popularity levels, which are least popular, below average, above average and most popular, based on their total review count.

The final optimization step involves evaluating product performance by location. Products were grouped according to their listed locations, with relevant attributes such as product price and quantity sold. For each location, the average product price and total quantity sold were computed. These figures were used to estimate market performance, calculated by multiplying the average price by the total quantity sold. Locations were then ranked from highest to lowest based on this performance indicator, enabling identification of regions with the strongest sales activity.

1.2 Objectives

The main objectives of this project are as follows:

- To develop a web crawler capable of extracting a minimum of 100,000 structured records from a targeted Malaysian e-commerce website.
- To apply high-performance computing techniques, including multithreading, multiprocessing and distributed processing, to optimise the efficiency and scalability of the web crawling and data processing systems.
- To implement ethical web scraping practices by respecting crawl delays and website usage policies.
- To conduct a comparative performance analysis of different data processing frameworks (Pandas, Polars and PySpark) based on the time consumed during data processing.
- To enhance students' technical proficiency, critical thinking in system optimization and collaborative skills in a diverse team environment.

1.3 Target website and data to be extracted

For this project, Lazada Malaysia (https://www.lazada.com.my/) was selected as the target website. Lazada is one of the leading e-commerce platforms in Southeast Asia, offering a wide range of products across multiple categories. The focus of the data extraction is on products under the "Women" category, which includes the following subcategories: Women's Fashion, Stationery, Mother and Baby, Home and Living, Health and Wellness and Beauty and Care. The fields extracted for each product are:

- **Product Name:** The title or description of the product as displayed on the website.
- **Location:** The seller's or product's listed location.
- **Quantity Sold:** The number of units sold, indicating the popularity of the product.
- **Price:** The listed selling price of the product.
- **Total reviews:** The total number of customer ratings received by the product.

Data scraping was carried out by applying a mix of Python libraries and tools such as BeautifulSoup, Selenium, Requests for complete data extraction. The stocks of data collected were then manipulated using pandas polars and PySpark, the processing time compared to evaluate performance enhancement in varied optimisation techniques.

2.0 System Design & Architecture

2.1 Description of architecture

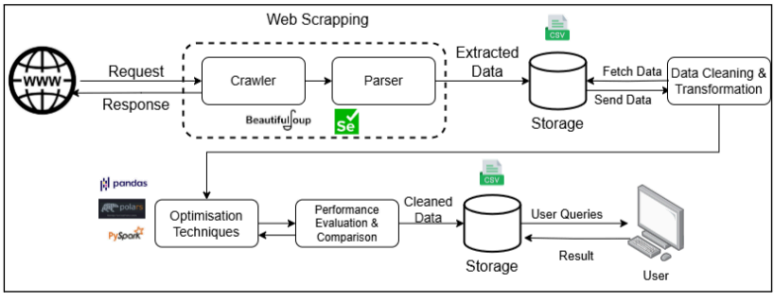


Figure 1: Web Crawler System Architecture

This project describes the design and the implementation of web crawler system for data extraction mechanisation, cleaning, optimisation and analysis. In this project data under use was acquired from Lazada Malaysia with the particular interest of the Women's category. The obtained dataset is termed "Women's Purchase Analysis".

The system starts when it sends a request to the target website and get the corresponding response.

- The Crawler component traverses the web pages in an organized manner, therefore, to gather pertinent data.
- The parser, by applying libraries, such as BeautifulSoup, and Selenium, retrieves structured data from the gathered content on the web.
- Then the extracted data is saved to a CSV file, for the subsequent process.

Following data extraction:

- A data cleaning and transformation is performed in retrieving the extracted data, removing inconsistencies, missing values and standardising the dataset to maintain quality and consistency.
- Cleaned data is stored in a structured form, separately.

For performance enhancement:

- Three libraries of Pandas, Polars and PySpark are applied so as to enhance the speed in processing data.
- A performance analysis and comparison is presented to run and compare the effectiveness of these selected optimisation methods.

Last:

- Cleaned and optimised data is provided for user queries.
- The users can interact interacting with the system by sending a query and the system will generate the requested analysis results based on the processed data set.

This system guarantees full workflow in data acquisition to effective data retrieval providing full analysis of women's purchase trends on Lazada.

2.2 Tools and frameworks used

The following tools and framework were used during the project:

- **BeautifulSoup**: A Python library applied for parsing HTML and XML documents. It made it easier to retrieve the targeted information from the Lazada web pages as the web scraping phase.
- **Selenium**: A web automation tool which was used to communicate with dynamic pages and deal with content that needed users to scroll and click to completely load before extraction.
- **Python**: The foremost language that was used for implementation of the web scraping, data cleaning and data analysis function.
- **Pandas**: A Python library for manipulating and examining data. It was used for cleaning, transformation and processing of the extracted data effectively.
- **Polars**: A fast Rust implementation of a Dataframe library, an alternative to Pandas for Dataframe manipulation for when dealing with bigger datasets.
- **PySpark**: The Python API for Apache Spark, developed to optimize processing of larger datasets, and distributed data operations to enhance performance and scalability produces considerably fewer errors.

2.3 Roles of team members

Bernice Lim Jing Xuan	● Project planning and management ● Developed optimization code using pandas
Kek Jesslyn	● Web crawling and scraping ● Developed scrapers using BeautifulSoup and Selenium ● Conducted performance comparison testing using laptop
Navachander Navasantar	● Developed optimization code using PySpark ● Assisted in report documentation
Tan Jun Yuan	● Developed optimization code using polars ● Assisted in report work planning and documentation

Table 1: Roles of Team Members

3.0 Data Collection

3.1 Crawling method

The web scraping script functioned with Python tools that used Selenium to control browsers and BeautifulSoup to extract HTML data. Our scraping solution undertook multiple page operations in female-oriented Lazada product sections including fashion and skincare along with wellness and home and baby and stationery selections.

- The scraping system detected page count through pagination components before it automatically visited all accessible pages in each category.
- The scraper included rate-limiting functions that simulated human behaviors by introducing random sleep time between 2.5 and 5 seconds when triggering actions such as page loading or "next" button clicks.
- The scraper relies on async handling while using Selenium's WebDriverWait to monitor page element load times as a method to avoid incomplete content during processing. Manual detection occurs when CAPTCHAs appear since a temporary stop occurs for user confirmation.

3.2 Number of records collected

A total of 115090 records were obtained from 36 distinct URLs that covered multiple pricing segments and sections including Women's Fashion, Beauty Skincare, Health Wellness, Home Living, Home Appliances, Mother & Baby and Stationery. Each record consists of:

- Product name
- Price
- Seller location
- Quantity sold
- Total reviews

3.3 Ethical considerations

The data scraping operation abided by all ethical standards throughout the process.

- Research and academic needs formed the sole basis for the scraping activities.
- The company Lazada has not declared any restrictions on data scraping for our collected data so we protected their server by delaying requests while managing concurrent requests at a moderate level.
- The researchers manually dealt with CAPTCHA interruptions to prevent any security bypassing situations.
- Our data scraping operations did not involve any wishes of personalized user information.
- We would either have accessed the Lazada API or attempted to obtain permission through clear terms if the platform provided either option.

4.0 Data Processing

4.1 Cleaning methods

The following cleaning techniques were applied to ensure the accuracy, completeness and consistency of the raw data:

- Merged seven individual data frames into a single unified dataframe.
- Verified the combined dataset by checking the total number of rows and columns.
- Removed irrelevant text and symbols from fields such as "Quantity Sold", including handling shorthand notations (example: converting "1.3K" to 1300).
- Converted non-numeric fields into appropriate numeric types where necessary, using error coercion to handle invalid values.
- Identified and handled missing data:
    - Replaced missing numeric entries with 0.
    - Replaced missing string entries with "N/A".
- Detected and removed duplicate rows to prevent redundancy and ensure data integrity.

4.2 Data structure

The final cleaned dataset was stored in the CSV (Comma-Separated Values) format, providing a lightweight and widely supported structure suitable for data processing and analysis tasks.

4.3 Transformation and formatting

After cleaning, the following transformations and formatting operations were performed to prepare the data for analysis and optimisation:

- Standardised all string fields by converting text to uppercase for consistency across entries.
- Ensured numeric fields, such as "Quantity Sold" and "Number of Ratings," were properly cast to integer (Int64) data types.
- Formatted the dataset to maintain uniform data types across all columns, facilitating smoother downstream processing.
- Structured the data to eliminate inconsistencies, enabling compatibility with optimisation libraries such as pandas, polars and PySpark.

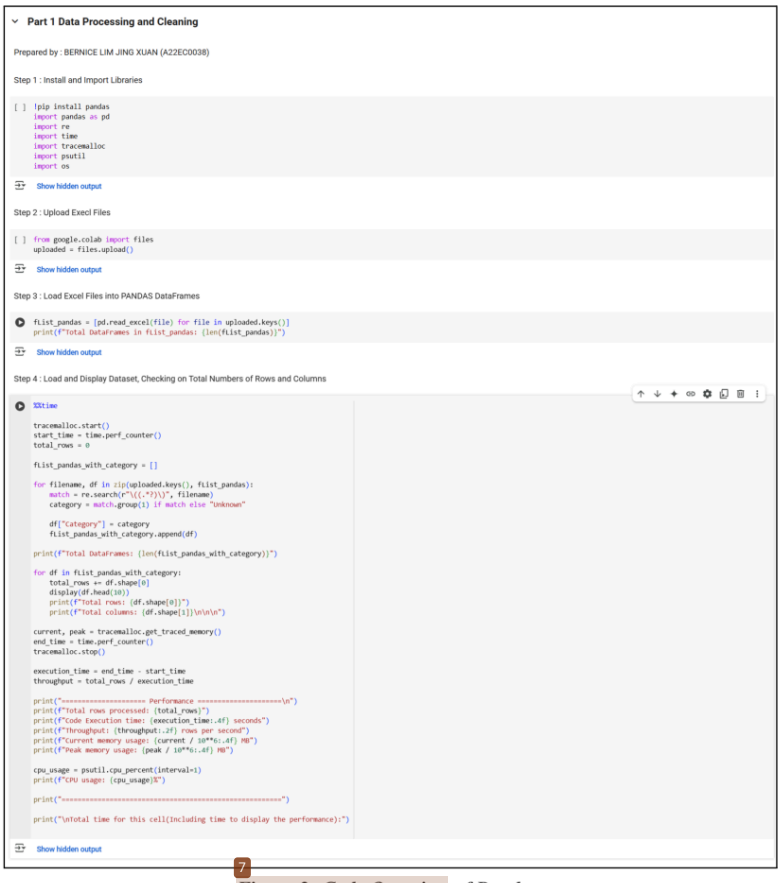
5.0 Optimization Techniques

5.1 Methods used: multithreading, multiprocessing, Spark, etc.

To optimize the data processing step, three python libraries were utilized; Pandas, polars and Pyspark libraries. Pandas was initially used as a benchmark because it is straightforward, and has many powerful data manipulation abilities. However, as Pandas functions in a single threaded approach it was found to have limitations with large datasets. To ensure a faster process, Polars was introduced. Polars supports multithreading and lazy evaluation which means that operations can compute across several CPU cores at once, and consequently are faster than Pandas.

Finally, for distributed processing an experimental and open source framework written in Python that runs on top of Apache Spark called PySpark was used. PySpark allows processing data at parallel across several cores or machines hence amazingly suitable for very large datasets. Its distributed architecture and native optimisation properties enabled large scale data transformation to work efficiently. Using Pandas (single-threaded), and the multithreaded variant of Polars and PySpark (distributed), the project evaluated the performance and scalability of various optimisation methods for processing data web-scraped.

5.2 Code overview or pseudocode of techniques applied



Part 1 Data Processing and Cleaning

Prepared by : BERNICE LIM JING XUAN (A22EC0038)

Step 1 : Install and Import Libraries

```
[ ] !pip install pandas
    import pandas as pd
    import re
    import time
    import tracemalloc
    import psutil
    import os
```

Show hidden output

Step 2 : Upload Execl Files

```
[ ] from google.colab import files
    uploaded = files.upload()
```

Show hidden output

Step 3 : Load Excel Files into PANDAS DataFrames

```
fList_pandas = [pd.read_excel(file) for file in uploaded.keys()]
print(f"Total DataFrames in fList_pandas: {len(fList_pandas)}")
```

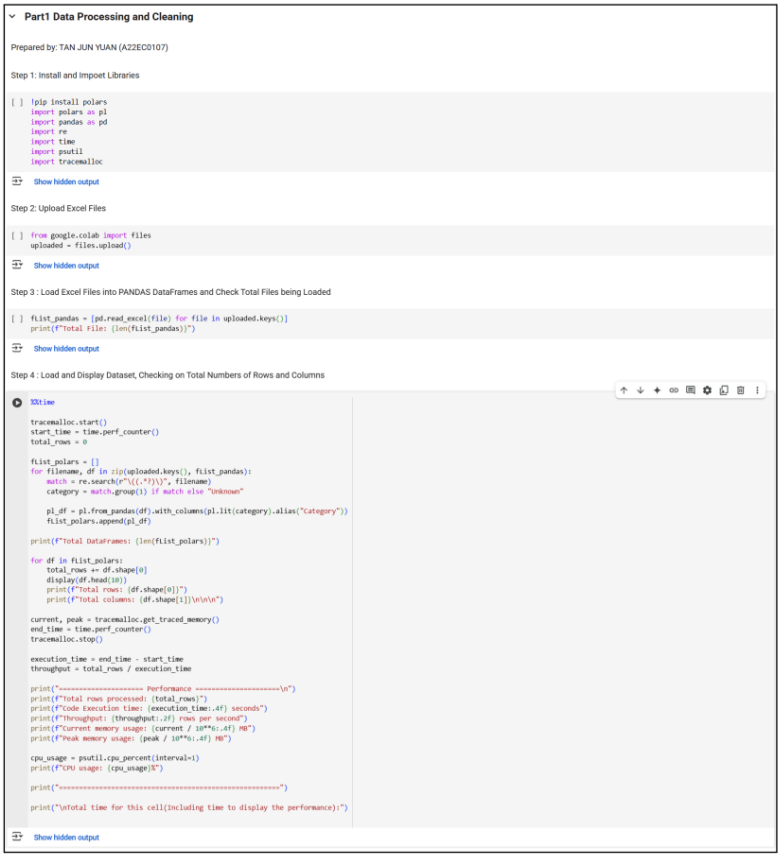
Show hidden output

Step 4 : Load and Display Dataset, Checking on Total Numbers of Rows and Columns

```
%%time

tracemalloc.start()
start_time = time.perf_counter()
total_rows = 0

fList_pandas_with_category = []

for filename, df in zip(uploaded.keys(), fList_pandas):
    match = re.search(r"\((.*?)\)", filename)
    category = match.group(1) if match else "Unknown"

    df["Category"] = category
    fList_pandas_with_category.append(df)

print(f"Total DataFrames: {len(fList_pandas_with_category)}")

for df in fList_pandas_with_category:
    total_rows += df.shape[0]
    display(df.head(10))
    print(f"Total rows: {df.shape[0]}")
    print(f"Total columns: {df.shape[1]}\n\n\n")

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time
throughput = total_rows / execution_time

print("==================== Performance ====================\n")
print(f"Total rows processed: {total_rows}")
print(f"Code Execution time: {execution_time:.4f} seconds")
print(f"Throughput: {throughput:.2f} rows per second")
print(f"Current memory usage: {current / 10**6:.4f} MB")
print(f"Peak memory usage: {peak / 10**6:.4f} MB")

cpu_usage = psutil.cpu_percent(interval=1)
print(f"CPU usage: {cpu_usage}%")

print("=====================================================")

print("\nTotal time for this cell(Including time to display the performance):")
```

Show hidden output

Figure 2: Code Overview of Pandas



Part1 Data Processing and Cleaning

Prepared by: TAN JUN YUAN (A22EC0107)

Step 1: Install and Impoet Libraries

```
[ ] !pip install polars
    import polars as pl
    import pandas as pd
    import re
    import time
    import psutil
    import tracemalloc
```

Show hidden output

Step 2: Upload Excel Files

```
[ ] from google.colab import files
    uploaded = files.upload()
```

Show hidden output

Step 3 : Load Excel Files into PANDAS DataFrames and Check Total Files being Loaded

```
[ ] flist_pandas = [pd.read_excel(file) for file in uploaded.keys()]
    print(f"Total File: {len(flist_pandas)}")
```

Show hidden output

Step 4 : Load and Display Dataset, Checking on Total Numbers of Rows and Columns

```
%%time

tracemalloc.start()
start_time = time.perf_counter()
total_rows = 0

flist_polars = []
for filename, df in zip(uploaded.keys(), flist_pandas):
    match = re.search(r"\((.*?)\)", filename)
    category = match.group(1) if match else "Unknown"

    pl_df = pl.from_pandas(df).with_columns(pl.lit(category).alias("Category"))
    flist_polars.append(pl_df)

print(f"Total DataFrames: {len(flist_polars)}")

for df in flist_polars:
    total_rows += df.shape[0]
    display(df.head(10))
    print(f"Total rows: {df.shape[0]}")
    print(f"Total columns: {df.shape[1]}\n\n\n")

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time
throughput = total_rows / execution_time

print("==================== Performance ====================\n")
print(f"Total rows processed: {total_rows}")
print(f"Code Execution time: {execution_time:.4f} seconds")
print(f"Throughput: {throughput:.2f} rows per second")
print(f"Current memory usage: {current / 10**6:.4f} MB")
print(f"Peak memory usage: {peak / 10**6:.4f} MB")

cpu_usage = psutil.cpu_percent(interval=1)
print(f"CPU usage: {cpu_usage}%")

print("=====================================================")

print("\nTotal time for this cell(Including time to display the performance):")
```

Show hidden output

Figure 3: Code Overview of Polars

Part1 Data Processing and Cleaning

Prepared by:

Step 1: Install and Impoet Libraries

```
from pyspark.sql import SparkSession
from pyspark.sql.types import StringType, NumericType
from pyspark.sql.functions import col, upper, regexp_replace, when, isnan, count, lit, mean, sum as spark_sum, avg
from pyspark.sql import functions as F
from functools import reduce
import pandas as pd
import re
import time
import tracemalloc
import psutil
import os
import shutil

spark = SparkSession.builder.appName("ExcelProcessing").getOrCreate()
```

Step 2: Upload Excel Files

```
from google.colab import files
uploaded = files.upload()
```

Show hidden output

Step 3 : Load Excel Files into PANDAS DataFrames and Check Total Files being Loaded

```
flist_pandas = [pd.read_excel(file) for file in uploaded.keys()]
print(f"Total DataFrames in flist_pandas: {len(flist_pandas)}")
```

Show hidden output

Step 4 : Load and Display Dataset, Checking on Total Numbers of Rows and Columns

```
%%time

tracemalloc.start()
start_time = time.perf_counter()
total_rows = 0

placeholders = ["N/A", "NA", "null", "", "--", "-", "n/a", "nan", "NaN"]
flist_spark = []

for filename, df in zip(uploaded.keys(), [pd.read_excel(file) for file in uploaded.keys()]):
    match = re.search(r"\([^()]+)\)", filename)
    category = match.group(1).strip() if match else "Unknown"

    df_cleaned = df.replace(placeholders, None)

    spark_df = spark.createDataFrame(df_cleaned)

    spark_df = spark_df.select([
        when(col(c).isin(placeholders), None).otherwise(col(c)).alias(c)
        if spark_df.schema[c].dataType.simpleString() == "string" else col(c)
        for c in spark_df.columns
    ])

    spark_df = spark_df.withColumn("Category", lit(category))

    flist_spark.append(spark_df)

print(f"Total DataFrames in flist_spark: {len(flist_spark)}")

for df in flist_spark:
    display(df.limit(10).toPandas())
    row_count = df.count()
    col_count = len(df.columns)
    total_rows += row_count
    print(f"Total rows: {row_count}")
    print(f"Total columns: {col_count}\n")

current, peak = tracemalloc.get_traced_memory()
end_time = time.perf_counter()
tracemalloc.stop()

execution_time = end_time - start_time
throughput = total_rows / execution_time

print("========== Performance ==========\n")
print(f"Total rows processed: {total_rows}")
print(f"Code Execution time: {execution_time:.4f} seconds")
print(f"Throughput: {throughput:.2f} rows per second")
print(f"Current memory usage: {current / 10**6:.4f} MB")
print(f"Peak memory usage: {peak / 10**6:.4f} MB")

cpu_usage = psutil.cpu_percent(interval=1)
print(f"CPU usage: {cpu_usage}%")

print("\nTotal time for this cell (Including time to display the performance):")
```

Show hidden output

Figure 4: Code Overview of PySpark

6.0 Performance Evaluation

6.1 Before vs after optimization

Initially, all data processing tasks were carried out using Pandas. While Pandas performed well for small datasets, it became noticeably slower as the data size increased, especially with operations like filtering, aggregation and joining. This led to delays that affected the overall workflow efficiency.

To improve performance, two faster alternatives were introduced, which are Polars and PySpark. Polars with its multi-threaded Rust backend, offered significant speed improvements for in-memory operations, while PySpark provided better handling for larger datasets through distributed processing even in local mode. After optimisation, execution times were greatly reduced, with Polars offering the fastest performance and PySpark also outperforming Pandas, although with a slight overhead due to its distributed nature. Overall, the optimisation resulted in a much faster and more scalable data processing pipeline.

6.2 Comparison of Code Execution Time, Peak Memory Usage, CPU usage and Throughput

Operation	Aspects	Comparisons		
		Pandas	Polars	Pyspark
Dataset Loading and Display	Code Execution Time (s)	0.6728	0.3893	113.0075
	Peak Memory Usage (MB)	1.4805	2.9626	34.7047
	CPU Usage (%)	2.5	2.0	84.8
	Throughput (rows/s)	171056.76	295636.96	1018.43
Dataset Integration	Code Execution Time (s)	0.0907	0.0207	4.8765
	Peak Memory Usage (MB)	5.6588	0.0419	0.1984
	CPU Usage (%)	3.5	2.5	12.6
	Throughput (rows/s)	1269298.78	5564845.61	23601.16
Standardization of String Data	Code Execution Time (s)	0.6550	0.0447	4.2237
	Peak Memory Usage (MB)	50.4867	0.0408	0.1824
	CPU Usage (%)	3.5	2.0	61.0
	Throughput (rows/s)	175721.43	2574996.64	27248.92
Convert “Total Reviews” to correct data type	Code Execution Time (s)	2.8809	0.0260	5.1355
	Peak Memory Usage (MB)	12.7351	0.0420	0.1679

(int)	CPU Usage (%)	4.0	2.5	21.6
	Throughput (rows/s)	39949.87	4434821.69	22410.57
Convert "Quantity Sold" to correct data type (int)	Code Execution Time (s)	3.9448	0.0843	4.1328
	Peak Memory Usage (MB)	19.3337	0.3887	0.1938
	CPU Usage (%)	25.6	2.0	6.0
	Throughput (rows/s)	29174.87	1365230.27	27847.80
Check and Handle Missing Values	Code Execution Time (s)	0.3365	0.0300	18.2981
	Peak Memory Usage (MB)	19.8926	0.0498	0.2372
	CPU Usage (%)	3.5	2.0	12.5
	Throughput (rows/s)	341984.53	3836148.05	6289.71
Check and Handle Duplicates	Code Execution Time (s)	0.6107	0.0859	65.1819
	Peak Memory Usage (MB)	20.1169	0.0745	2.3469
	CPU Usage (%)	3.5	2.5	13.6
	Throughput (rows/s)	188440.81	1339841.70	1742.75
Export cleaned dataset file for optimization	Code Execution Time (s)	4.50939	0.0592	10.6576
	Peak Memory Usage (MB)	3.8787	0.0953	0.0880
	CPU Usage (%)	3.5	6.0	75.1
	Throughput (rows/s)	25191.72	1918315.45	10658.73

Table 2: Comparison between Data Processing and Cleaning Techniques

Operation	Aspects	Comparisons		
		Pandas	Polars	Pyspark
Grouping Products into 4 categories based on price (Budget Friendly, Affordable, Mid-Range and Premium Price)	Code Execution Time (s)	0.6036	0.5189	22.6163
	Peak Memory Usage (MB)	19.2549	0.1225	0.3516
	CPU Usage (%)	88.6	90.0	5.5
	Throughput (rows/s)	188187.82	218918.57	5022.74
Grouping Products into 4 categories based on 'Total Reviews' (Least, Below Average, Above Average and Most Popular)	Code Execution Time (s)	0.9005	0.2981	13.5979
	Peak Memory Usage (MB)	15.0830	0.0664	0.3121
	CPU Usage (%)	86.9	100.0	19.0
	Throughput (rows/s)	126141.24	381122.98	8353.95
Evaluate and Rank Market Performance based on "Quantity Sold" for each "Location"	Code Execution Time (s)	2.1653	0.1121	1.8099
	Peak Memory Usage (MB)	6.2297	0.0222	0.1632
	CPU Usage (%)	95.5	100.0	56.9
	Throughput (rows/s)	52461.36	1013159.21	62763.71

Table 3: Comparison between Data Optimization Techniques

Conclusion:

Overall performance between the libraries: **Polars > Pandas > PySpark**

- **Polars:** Polars delivered superior performance than PySpark on big datasets thanks to its Rust-based foundation coupled with threading capabilities and columnar data structure.
- **Pandas:** The system efficiency of Pandas was high but it utilised more system memory and took longer to process data.
- **PySpark:** PySpark registered the slowest performance since it required high resource allocation latency combined with substantial initialisation overhead.

6.3 Charts and graphs

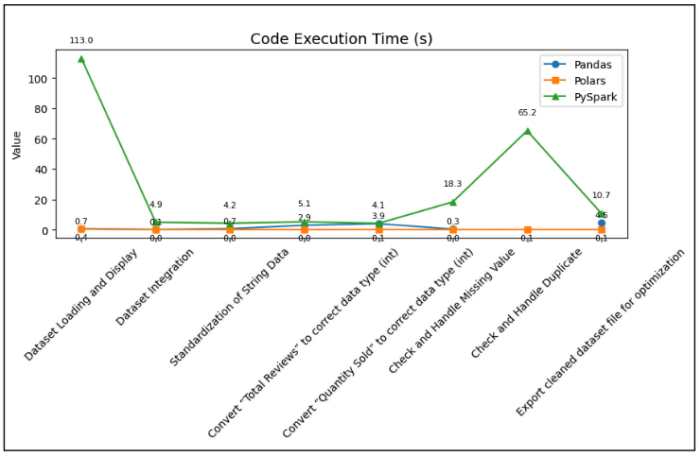


Figure 5: Line graph for Code Execution Time (s)

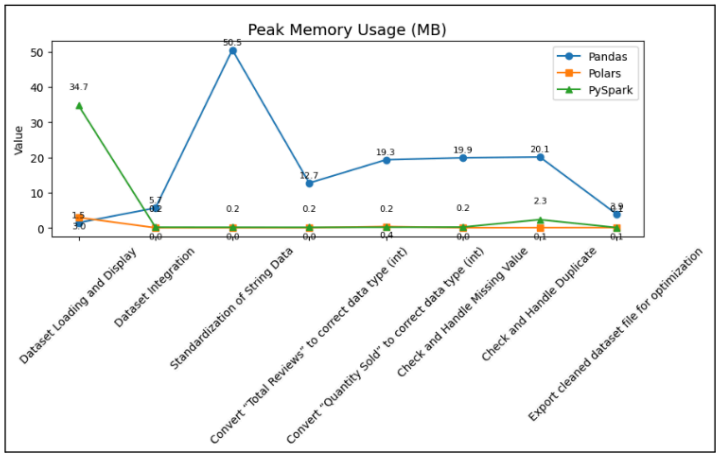


Figure 6: Line graph for Peak Memory Usage (MB)

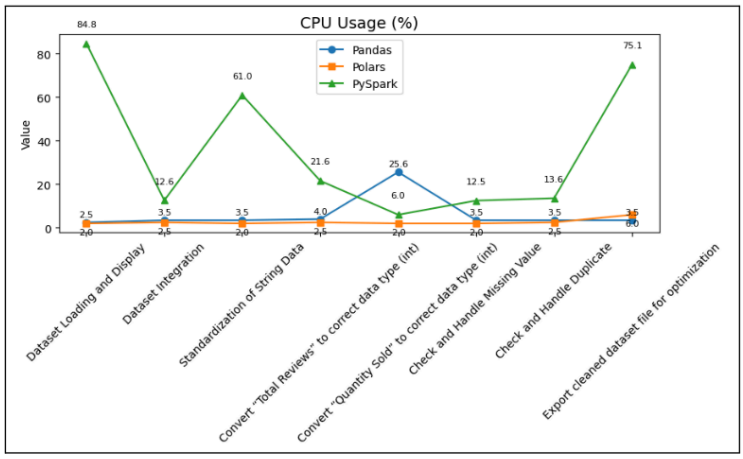


Figure 7: Line graph for CPU Usage (%)

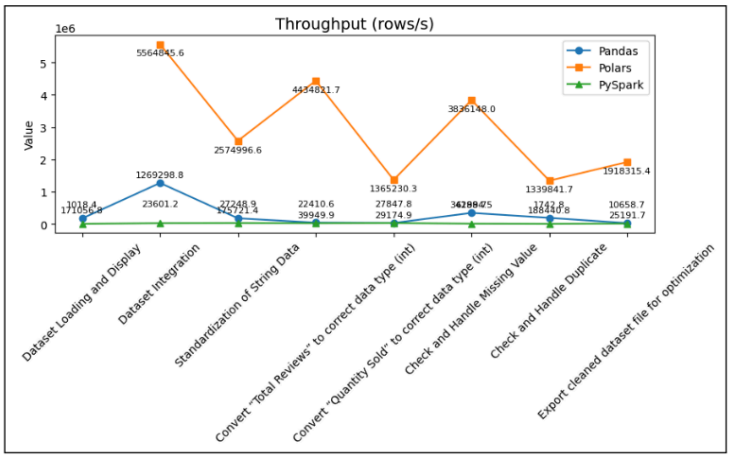


Figure 8: Line graph for Throughput (rows/s)

7.0 Challenges & Limitations

7.1 What didn't go as planned

Initially, the project intended to use only Requests and BeautifulSoup for web scraping. However, Lazada's dynamic JavaScript-driven content made it impossible to retrieve the necessary information with these tools alone. Therefore, Selenium was incorporated to handle the dynamic loading of data. However, using Selenium brought new issues, particularly the need for ChromeDriver installation. Since Google Colab could not successfully run ChromeDriver, the team had to shift the environment to Visual Studio Code on local machines. This transition increased the setup time and complexity.

Another major challenge was Lazada's frequent CAPTCHA verifications. The team had to manually solve CAPTCHAs during the scraping process, which greatly slowed down data collection. Scraping had to be done during weekends, requiring long hours in front of the computer to monitor and complete the verification steps.

During the data processing stage, differences in library capabilities also posed problems. For example, Polars could not directly read Excel files. To resolve this, the data was first read into Pandas and then converted into a Polars DataFrame, adding extra steps and minor inefficiencies to the workflow.

7.2 Any limitations of your solution

Despite overcoming many challenges, some limitations remained. The most significant was the reliance on manual CAPTCHA handling, which prevented full automation and reduced scraping efficiency. Running Selenium locally on Visual Studio Code also restricted collaboration, as each member needed to configure their own environment separately. This limited the flexibility that cloud platforms like Colab would have provided.

In addition, extra conversion steps between data formats affected the purity of the performance comparisons between Pandas, Polars and PySpark. Although functional, it introduced slight variations in the benchmarking results.

Finally, the need for manual oversight and longer scraping times meant that the project could not easily scale to even larger datasets or implement more advanced scraping strategies like headless browsing or automated CAPTCHA solving.

8.0 Conclusion & Future Work

8.1 Summary of findings

This project sought to improve high-efficiency data processing for massive web crawling of Lazada Malaysia's Women category. Our group collected over 115,000 product listings by utilising the tools named Selenium and BeautifulSoup to navigate the site and pull out information such as product titles, price, number sold, seller locations, and customer reviews. We filtered and arranged the data once it was gathered, prepared for analysis.

We then contrasted and examined three distinct libraries to determine which one was most appropriate to process large-scale data effectively:

- Pandas, though easy to use and effective for small to midsize datasets, showed worsening performance and more memory usage with our large dataset and thus turned out to be less effective for large-scale data processing tasks.
- Polars, with a Rust backend and lazy evaluation approach, exhibited excellent performance gains in speed and memory consumption. It offered a perfect harmony between usability and functionality for moderately sized datasets.
- PySpark was typified by scalability, distributed processing, and appropriateness for large-scale data projects. Still, its configurational complexity and associated overhead rendered it less suitable for small or medium-size projects.

Overall, this project provided our group with hands-on practice, from data collections and cleaning to performance testing of modern data processing tools. Most significantly, we learned how to assess and choose suitable technologies depending on certain requirements such as dataset size, processing time, and memory consumption. This project not only improved our technical expertise in Python and high performance libraries but also improved our critical thinking, collaboration, and problem solving skills, equipping us for upcoming projects in data engineering and data analysis.

8.2 What could be improved

While the current implementation successfully achieved its primary goals, there are several meaningful ways the project can be expanded and enhanced in the future:

1. Automate CAPTCHA Handling for seamless handling

Manual CAPTCHA solving during scraping delayed the data collection process. 2Captcha browser extension allows skipping reCAPTCHAs. This extension automates the process of solving reCAPTCHA, making it easier and faster for users to bypass these verifications [1]. Incorporating automated services such as 2Captcha can automate the process with machine learning powered CAPTCHA solvers.

2. Make the most of GPU Powered Libraries for better performance

Libraries such as RAPIDS cuDF by NVIDIA can leverage GPUs to greatly accelerate data processing operations. For instance, cuDF accelerates pandas with zero code changes and brings greatly improved performance [2]. This is particularly useful when working with large numeric datasets.

3. Use machine learning for further analysis

The structured data can be utilised to implement machine learning algorithms for customer segmentation. Machine learning methodologies are a great tool for analyzing customer data and finding insights and patterns. Artificially intelligent models are powerful tools for decision-makers. They can precisely identify customer segments, which is much harder to do manually or with conventional analytical methods [3]. For instance, K-Means is efficient machine learning when it comes to solving data cluster problems.

By adopting these improvements, the project has the potential to evolve into a robust, scalable and intelligent data pipeline capable of supporting practical, data-driven decisions in the e-commerce landscape.

References

[1] Captcha Solver: reCAPTCHA solver and captcha solving service. Bypass captchas using the best auto captcha solver online API - 2Captcha. (2025). 2captcha.com. https://2captcha.com/

[2] Open GPU Data Science. (n.d.). RAPIDS. https://rapids.ai/

[3] Kumar, D. (2021, June 18). Implementing Customer Segmentation Using Machine Learning [Beginners Guide]. Neptune.ai. https://neptune.ai/blog/customer-segmentation-using-machine-learning

Appendices

Figure 9: Group photo

Sample code snippets

1. **Web Scraping:** Codes
2. **Pandas:**
   - Part 1
   - Part 2
3. **Polars:**
   - Part 1
   - Part 2
4. **PySpark:**
   - Part 1
   - Part 2
5. **Graph:** Codes

Screenshots of output

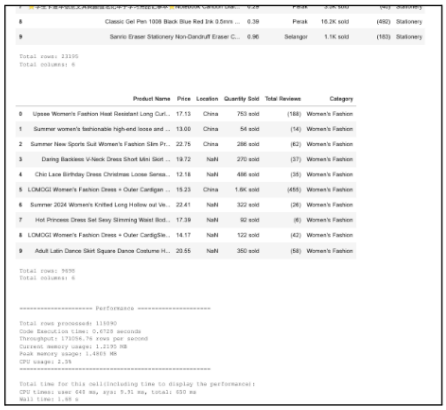


Figure 9: Output Screenshot of Dataset Loading and Display by using Pandas

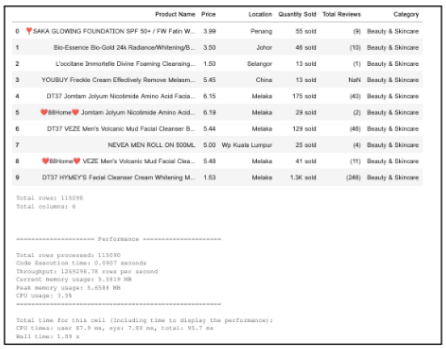
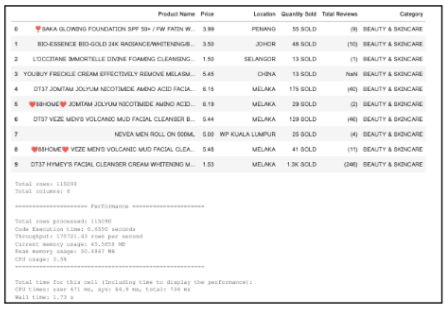


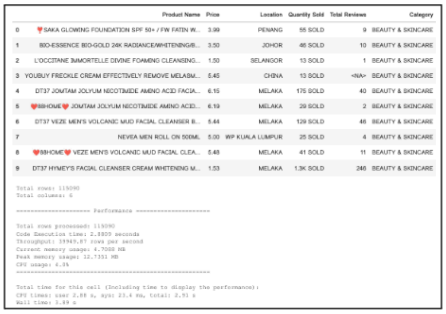
Figure 10: Output Screenshot of Dataset Integration by using Pandas



	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	❤️SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	PENANG	55 SOLD	(9)	BEAUTY & SKINCARE
1	BIO-ESSENCE BIO-GOLD 24K RADIANCE/WHITENING/B...	3.50	JOHOR	46 SOLD	(10)	BEAUTY & SKINCARE
2	L'OCCITANE IMMORTELLE DIVINE FOAMING CLEANSING...	1.50	SELANGOR	13 SOLD	(1)	BEAUTY & SKINCARE
3	YOUBUY FRECKLE CREAM EFFECTIVELY REMOVE MELASM...	5.45	CHINA	13 SOLD	NaN	BEAUTY & SKINCARE
4	DT37 JOMTAM JOLYUM NICOTIMIDE AMINO ACID FACIA...	6.15	MELAKA	175 SOLD	(40)	BEAUTY & SKINCARE
5	❤️BBHOME❤️ JOMTAM JOLYUM NICOTIMIDE AMINO ACID...	6.19	MELAKA	29 SOLD	(2)	BEAUTY & SKINCARE
6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129 SOLD	(46)	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 50ML	5.00	WP KUALA LUMPUR	25 SOLD	(4)	BEAUTY & SKINCARE
8	❤️BBHOME❤️ VEZE MEN'S VOLCANIC MUD FACIAL CLEA...	5.48	MELAKA	41 SOLD	(11)	BEAUTY & SKINCARE
9	DT37 HYMEY'S FACIAL CLEANSER CREAM WHITENING M...	1.53	MELAKA	1.3K SOLD	(246)	BEAUTY & SKINCARE

```
Total rows: 115090
Total columns: 6

========== Performance ==========

Total rows processed: 115090
Code Execution time: 0.6550 seconds
Throughput: 175721.65 rows per second
Current memory usage: 45.5858 MB
Peak memory usage: 50.4847 MB
CPU usage: 3.5%
--------------------------------------------------

Total time for this cell (Including time to display the performance):
CPU times: user 671 ms, sys: 84.9 ms, total: 756 ms
Wall time: 1.73 s
```

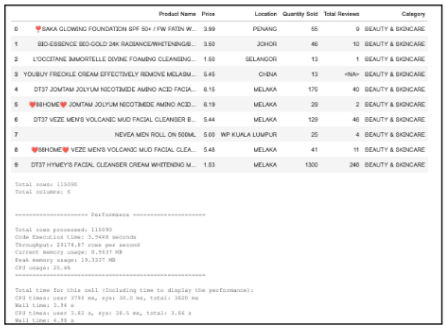
Figure 11: Output Screenshot of Standardization of String Data by using Pandas



	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	❤️SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	PENANG	55 SOLD	9	BEAUTY & SKINCARE
1	BIO-ESSENCE BIO-GOLD 24K RADIANCE/WHITENING/B...	3.50	JOHOR	46 SOLD	10	BEAUTY & SKINCARE
2	L'OCCITANE IMMORTELLE DIVINE FOAMING CLEANSING...	1.50	SELANGOR	13 SOLD	1	BEAUTY & SKINCARE
3	YOUBUY FRECKLE CREAM EFFECTIVELY REMOVE MELASM...	5.45	CHINA	13 SOLD	<NA>	BEAUTY & SKINCARE
4	DT37 JOMTAM JOLYUM NICOTIMIDE AMINO ACID FACIA...	6.15	MELAKA	175 SOLD	40	BEAUTY & SKINCARE
5	❤️BBHOME❤️ JOMTAM JOLYUM NICOTIMIDE AMINO ACID...	6.19	MELAKA	29 SOLD	2	BEAUTY & SKINCARE
6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129 SOLD	46	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 50ML	5.00	WP KUALA LUMPUR	25 SOLD	4	BEAUTY & SKINCARE
8	❤️BBHOME❤️ VEZE MEN'S VOLCANIC MUD FACIAL CLEA...	5.48	MELAKA	41 SOLD	11	BEAUTY & SKINCARE
9	DT37 HYMEY'S FACIAL CLEANSER CREAM WHITENING M...	1.53	MELAKA	1.3K SOLD	246	BEAUTY & SKINCARE

```
Total rows: 115090
Total columns: 6

========== Performance ==========

Total rows processed: 115090
Code Execution time: 2.8929 seconds
Throughput: 39949.87 rows per second
Current memory usage: 4.7088 MB
Peak memory usage: 12.7351 MB
CPU usage: 6.8%
--------------------------------------------------

Total time for this cell (Including time to display the performance):
CPU times: user 2.88 s, sys: 23.4 ms, total: 2.91 s
Wall time: 3.89 s
```

Figure 12: Output Screenshot of Converting "Total Reviews" to int Data Type by using Pandas



	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	❤️SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	PENANG	55	9	BEAUTY & SKINCARE
1	BIO-ESSENCE BIO-GOLD 24K RADIANCE/WHITENING/B...	3.50	JOHOR	46	10	BEAUTY & SKINCARE
2	L'OCCITANE IMMORTELLE DIVINE FOAMING CLEANSING...	1.50	SELANGOR	13	1	BEAUTY & SKINCARE
3	YOUBUY FRECKLE CREAM EFFECTIVELY REMOVE MELASM...	5.45	CHINA	13	<NA>	BEAUTY & SKINCARE
4	DT37 JOMTAM JOLYUM NICOTIMIDE AMINO ACID FACIA...	6.15	MELAKA	175	40	BEAUTY & SKINCARE
5	❤️BBHOME❤️ JOMTAM JOLYUM NICOTIMIDE AMINO ACID...	6.19	MELAKA	29	2	BEAUTY & SKINCARE
6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129	46	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 50ML	5.00	WP KUALA LUMPUR	25	4	BEAUTY & SKINCARE
8	❤️BBHOME❤️ VEZE MEN'S VOLCANIC MUD FACIAL CLEA...	5.48	MELAKA	41	11	BEAUTY & SKINCARE
9	DT37 HYMEY'S FACIAL CLEANSER CREAM WHITENING M...	1.53	MELAKA	1300	246	BEAUTY & SKINCARE

```
Total rows: 115090
Total columns: 6

========== Performance ==========

Total rows processed: 115090
Code Execution time: 2.7448 seconds
Throughput: 29174.67 rows per second
Current memory usage: 8.7637 MB
Peak memory usage: 19.3337 MB
CPU usage: 25.4%
--------------------------------------------------

Total time for this cell (Including time to display the performance):
CPU times: user 3781 ms, sys: 30.0 ms, total: 3820 ms
Wall time: 3.94 s
CPU times: user 3.82 s, sys: 38.5 ms, total: 3.86 s
Wall time: 4.88 s
```

Figure 13: Output Screenshot of Converting "Quantity Sold" to int Data Type by using Pandas

```
9  DT37 HYMEY'S FACIAL CLEANSER CREAM WHITENING M...  1.53  MELAKA  1300  246  BEAUTY & SKINCARE

Before Handling Missing Values:
Product Name      2
Price             0
Location      13317
Quantity Sold 45566
Total Reviews 58956
Category          0
dtype: int64

After Handling Missing Values:
Product Name      0
Price             0
Location          0
Quantity Sold     0
Total Reviews     0
Category          0
dtype: int64

Finalised Dataset:
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	❤️SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	PENANG	55	9	BEAUTY & SKINCARE
1	BIO-ESSENCE BIO-GOLD 24K RADIANCE/WHITENING/B...	3.50	JOHOR	46	10	BEAUTY & SKINCARE
2	L'OCCITANE IMMORTELLE DIVINE FOAMING CLEANSING...	1.50	SELANGOR	13	1	BEAUTY & SKINCARE
3	YOUBUY FRECKLE CREAM EFFECTIVELY REMOVE MELASM...	5.45	CHINA	13	0	BEAUTY & SKINCARE
4	DT37 JOMTAM JOLYUM NICOTIMIDE AMINO ACID FACIA...	6.15	MELAKA	175	40	BEAUTY & SKINCARE
5	❤️BBHOME❤️ JOMTAM JOLYUM NICOTIMIDE AMINO ACID...	6.19	MELAKA	29	2	BEAUTY & SKINCARE
6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129	46	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 50ML	5.00	WP KUALA LUMPUR	25	4	BEAUTY & SKINCARE
8	❤️BBHOME❤️ VEZE MEN'S VOLCANIC MUD FACIAL CLEA...	5.48	MELAKA	41	11	BEAUTY & SKINCARE
9	DT37 HYMEY'S FACIAL CLEANSER CREAM WHITENING M...	1.53	MELAKA	1300	246	BEAUTY & SKINCARE

```
Total rows: 115090
Total columns: 6

========== Performance ==========

Total rows processed: 115090
Code Execution time: 0.3365 seconds
Throughput: 341984.53 rows per second
Current memory usage: 6.5401 MB
Peak memory usage: 15.8926 MB
CPU usage: 3.5%
--------------------------------------------------

Total time for this cell (Including time to display the performance):
CPU times: user 269 ms, sys: 67.5 ms, total: 337 ms
Wall time: 0.34 s
CPU times: user 338 ms, sys: 9.8 ms, total: 348 ms
Wall time: 1.34 s
```

Figure 14: Output Screenshot of Checking and Handling Missing Values by using Pandas

2854 rows × 6 columns

Total rows: 2954
Total columns: 6

After Handling Duplicates
Number of duplicate rows: 0

Product Name	Price	Location	Quantity Sold	Total Reviews	Category

Total rows: 0
Total columns: 6

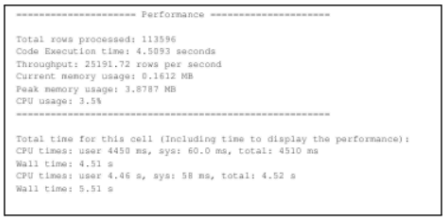
Finalised Dataset:

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	❤️SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	PENANG	55	9	BEAUTY & SKINCARE
1	SEO-ESSENCE SEO-GOLD 24K RADIANCE/WHITENING/B...	3.50	JOHOR	46	10	BEAUTY & SKINCARE
2	L'OCCITANE IMMORTELLE DIVINE FOAMING CLEANSING...	1.50	SELANGOR	13	1	BEAUTY & SKINCARE
3	YOUBUY FRECKLE CREAM EFFECTIVELY REMOVE MELASM...	5.40	CHINA	13	0	BEAUTY & SKINCARE
4	DT37 JOMTAM JOLYUN NEODTMIDE AMINO ACID FACIA...	6.15	MELAKA	175	40	BEAUTY & SKINCARE
5	❤️RIHOME❤️ JOMTAM JOLYUN NEODTMIDE AMINO ACID...	6.19	MELAKA	29	2	BEAUTY & SKINCARE
6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129	46	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 50ML	5.00	WP KUALA LUMPUR	25	4	BEAUTY & SKINCARE
8	❤️RIHOME❤️ VEZE MEN'S VOLCANIC MUD FACIAL CLEA...	5.48	MELAKA	61	11	BEAUTY & SKINCARE
9	DT37 HYMEYS FACIAL CLEANSER CREAM WHITENING M...	1.53	MELAKA	1300	246	BEAUTY & SKINCARE

Total rows: 113596
Total columns: 6

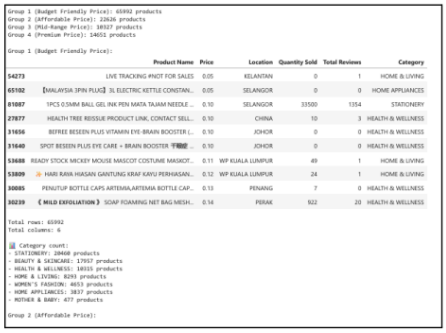
```
========== Performance ==========

Total rows processed: 113596
Code Execution time: 0.6107 seconds
Throughput: 186441.91 rows per second
Current memory usage: 9.2945 MB
Peak memory usage: 20.1149 MB
CPU usage: 3.5%
==================================================

Total time for this cell (Including time to display the performance):
CPU times: user 613 ms, sys: 0.0 ns, total: 613 ms
Wall time: 5.41 s
CPU times: user 635 ms, sys: 7.08 ms, total: 642 ms
Wall time: 1.42 s
```

Figure 15: Output Screenshot of Checking and Handling Duplicates by using Pandas



```
========== Performance ==========

Total rows processed: 113596
Code Execution time: 4.5093 seconds
Throughput: 25191.72 rows per second
Current memory usage: 0.1612 MB
Peak memory usage: 3.8787 MB
CPU usage: 3.5%
==================================================

Total time for this cell (Including time to display the performance):
CPU times: user 4450 ms, sys: 60.0 ms, total: 4510 ms
Wall time: 4.51 s
CPU times: user 4.46 s, sys: 58 ms, total: 4.52 s
Wall time: 5.51 s
```

Figure 16: Output Screenshot of Exporting Cleaned Dataset File by using Pandas

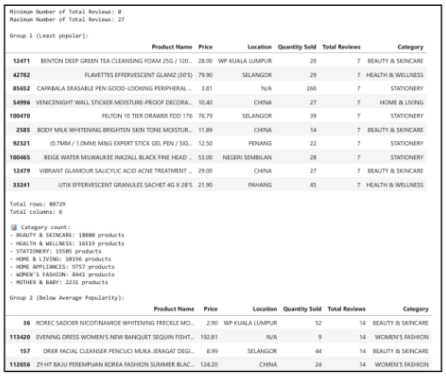


```
Group 1 (Budget Friendly Price): 65992 products
Group 2 (Affordable Price): 22626 products
Group 3 (Mid-Range Price): 10327 products
Group 4 (Premium Price): 14651 products

Group 1 (Budget Friendly Price):
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
54273	LIVE TRACKING #NOT FOR SALES	0.05	KELANTAN	0	1	HOME & LIVING
65102	【MALAYSIA 3PIN PLUG】3L ELECTRIC KETTLE CONSTAN...	0.05	SELANGOR	0	0	HOME APPLIANCES
81087	1PCS 0.5MM BALL GEL INK PEN MATA TAJAM NEEDLE ...	0.10	SELANGOR	33500	1354	STATIONERY
27877	HEALTH TREE REISSUE PRODUCT LINK, CONTACT SELL...	0.10	CHINA	10	3	HEALTH & WELLNESS
31656	BEFREE BESEEN PLUS VITAMIN EYE-BRAIN BOOSTER (...	0.10	JOHOR	0	0	HEALTH & WELLNESS
31640	SPOT BESEEN PLUS EYE CARE + BRAIN BOOSTER TREE	0.10	JOHOR	0	0	HEALTH & WELLNESS
53688	READY STOCK MICKEY MOUSE MASCOT COSTUME MASKOT...	0.11	WP KUALA LUMPUR	49	1	HOME & LIVING
53809	✨ HARI RAYA HIASAN GANTUNG KRAF-KAYU PERHIASAN...	0.12	WP KUALA LUMPUR	24	1	HOME & LIVING
30085	PENUTUP BOTTLE CAPS ARTEMIA,ARTEMIA BOTTLE CAP...	0.13	PENANG	7	0	HEALTH & WELLNESS
30239	《 MILD EXFOLIATION 》SOAP FOAMING NET BAG MESH...	0.14	PERAK	922	20	HEALTH & WELLNESS

```
Total rows: 65992
Total columns: 6

📊 Category count:
- STATIONERY: 20468 products
- BEAUTY & SKINCARE: 17957 products
- HEALTH & WELLNESS: 10315 products
- HOME & LIVING: 8293 products
- WOMEN'S FASHION: 4653 products
- HOME APPLIANCES: 3837 products
- MOTHER & BABY: 477 products

Group 2 (Affordable Price):
```

Figure 17: Output Screenshot of Grouping Product into 4 Categories based on "Price" by using Pandas



```
Minimum Number of Total Reviews: 0
Maximum Number of Total Reviews: 27

Group 1 (Least popular):
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
12471	BENTON DEEP GREEN TEA CLEANSING FOAM 25G / 120...	28.00	WP KUALA LUMPUR	20	7	BEAUTY & SKINCARE
42782	FLAVETTES EFFERVESCENT GLAMZ (10'S)	79.90	SELANGOR	29	7	HEALTH & WELLNESS
85652	CAPABALA ERASABLE PEN GOOD-LOOKING PERIPHERAL ...	3.81	N/A	266	7	STATIONERY
54996	VENCENIGHT WALL STICKER MOISTURE-PROOF DECORA...	10.40	CHINA	27	7	HOME & LIVING
100470	FELTON 10 TIER DRAWER FDD 176	76.79	SELANGOR	39	7	STATIONERY
2585	BODY MILK WHITENING BRIGHTEN SKIN TONE MOISTUR...	11.89	CHINA	14	7	BEAUTY & SKINCARE
92321	@.7MM / 1.0MM MSG EXPERT STICK GEL PEN / SIG...	12.50	PENANG	22	7	STATIONERY
100465	BEIGE WATER MUBAIUKIE INRZALL BLACK FINE HEAD ...	53.00	NEGERI SEMBILAN	28	7	STATIONERY
12479	VIBRANT GLAMOUR SALICYLIC ACID ACNE TREATMENT ...	29.00	CHINA	27	7	BEAUTY & SKINCARE
33241	UTIX EFFERVESCENT GRANULES SACHET 4G X 28'S	21.90	PAHANG	45	7	HEALTH & WELLNESS

```
Total rows: 80729
Total columns: 6

📊 Category count:
- BEAUTY & SKINCARE: 18080 products
- HEALTH & WELLNESS: 16519 products
- STATIONERY: 15505 products
- HOME & LIVING: 10196 products
- HOME APPLIANCES: 9757 products
- WOMEN'S FASHION: 8441 products
- MOTHER & BABY: 2231 products

Group 2 (Below Average Popularity):
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
36	ROREC SADOER NICOTINAMIDE WHITENING FRECKLE MO...	2.90	WP KUALA LUMPUR	52	14	BEAUTY & SKINCARE
113420	EVENING DRESS WOMEN'S NEW BANQUET SEQUIN FISHT...	192.81	N/A	9	14	WOMEN'S FASHION
157	DRKR FACIAL CLEANSER PENCUCI MUKA JERAGAT DEGI...	8.99	SELANGOR	44	14	BEAUTY & SKINCARE
112656	ZY-HT BAJU PEREMPUAN KOREA FASHION SUMMER BLAC...	124.20	CHINA	24	14	WOMEN'S FASHION

Figure 18: Output Screenshot of Grouping Products into 4 Categories based on "Total Reviews" by using Pandas

	Location	Total_Quantity_Sold	Average_Price	Market Performance
23	SELANGOR	10562173	100.849751	1.065193e+09
14	N/A	18524452	43.647313	8.085426e+08
4	CHINA	6341944	97.622898	6.191190e+08
29	WP KUALA LUMPUR	1227922	101.519881	1.246585e+08
8	JOHOR	1499826	78.073532	1.170967e+08
16	OVERSEAS	2083568	48.480572	1.010126e+08
18	PENANG	962481	100.382696	9.661644e+07
19	PERAK	1029478	80.170026	8.253328e+07
6	HONG KONG	376652	138.300619	5.209120e+07
9	KEDAH	595993	80.050847	4.770974e+07

```
==================== Performance ====================
Total rows processed: 113596
Code Execution time: 2.1653 seconds
Throughput: 52461.36 rows per second
Current memory usage: 0.0738 MB
Peak memory usage: 6.2297 MB
CPU usage: 95.5%
============================================================
CPU times: user 158 ms, sys: 4.42 ms, total: 163 ms
Wall time: 2.25 s
```

Figure 19: Output Screenshot of Evaluating and Ranking Market Performance based on "Quantity Sold" for each "Location" by using Pandas

shape: (10, 6)

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	str	str	str
"Upsee Women's Fashion Heat Res...	17.13	"China"	"753 sold"	"(188)"	"Women's Fashion"
"Summer women's fashionable hig...	13.0	"China"	"54 sold"	"(14)"	"Women's Fashion"
"Summer New Sports Suit Women's...	22.75	"China"	"286 sold"	"(62)"	"Women's Fashion"
"Daring Backless V-Neck Dress S...	19.72	null	"270 sold"	"(37)"	"Women's Fashion"
"Chic Lace Birthday Dress Chris...	12.18	null	"486 sold"	"(35)"	"Women's Fashion"
"LOMOGI Women's Fashion Dress +...	15.23	"China"	"1.6K sold"	"(455)"	"Women's Fashion"
"Summer 2024 Women's Knitted Lo...	22.41	null	"322 sold"	"(26)"	"Women's Fashion"
"Hot Princess Dress Set Sexy Sl...	17.39	null	"92 sold"	"(6)"	"Women's Fashion"
"LOMOGI Women's Fashion Dress +...	14.17	null	"122 sold"	"(42)"	"Women's Fashion"
"Adult Latin Dance Skirt Square...	20.55	null	"350 sold"	"(58)"	"Women's Fashion"

```
Total rows: 9698
Total columns: 6


==================== Performance ====================

Total rows processed: 115090
Code Execution time: 0.6578 seconds
Throughput: 174953.89 rows per second
Current memory usage: 2.7757 MB
Peak memory usage: 2.8360 MB
CPU usage: 57.8%
============================================================

Total time for this cell(Including time to display the performance):
CPU times: user 337 ms, sys: 42.8 ms, total: 380 ms
Wall time: 1.67 s
```

Figure 20: Output Screenshot of Dataset Loading and Display by using Polars

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	str	str	str
"❣️SAKA GLOWING FOUNDATION SPF ...	3.99	"Penang"	"55 sold"	"(9)"	"Beauty & Skincare"
"Bio-Essence Bio-Gold 24k Radi...	3.5	"Johor"	"46 sold"	"(10)"	"Beauty & Skincare"
"L'occitane Immortelle Divine F...	1.5	"Selangor"	"13 sold"	"(1)"	"Beauty & Skincare"
"YOUBUY Freckle Cream Effective...	5.45	"China"	"13 sold"	null	"Beauty & Skincare"
"DT37 Jomtam Jolyum Nicotimide ...	6.15	"Melaka"	"175 sold"	"(40)"	"Beauty & Skincare"
"❣️88Home❣️ Jomtam Jolyum Nicot...	6.19	"Melaka"	"29 sold"	"(2)"	"Beauty & Skincare"
"DT37 VEZE Men's Volcanic Mud F...	5.44	"Melaka"	"129 sold"	"(46)"	"Beauty & Skincare"
"NEVEA MEN ROLL ON 500ML"	5.0	"Wp Kuala Lumpur"	"25 sold"	"(4)"	"Beauty & Skincare"
"❣️88Home❣️ VEZE Men's Volcanic...	5.48	"Melaka"	"41 sold"	"(11)"	"Beauty & Skincare"
"DT37 HYMEY'S Facial Cleanser C...	1.53	"Melaka"	"1.3K sold"	"(246)"	"Beauty & Skincare"

```
Total rows: 115090
Total columns: 6


==================== Performance ====================

Total rows processed: 115090
Code Execution time: 0.1654 seconds
Throughput: 695962.63 rows per second
Current memory usage: 0.0284 MB
Peak memory usage: 0.0989 MB
CPU usage: 4.8%
============================================================

Total time for this cell(Including time to display the performance):
CPU times: user 95.5 ms, sys: 47.3 ms, total: 143 ms
Wall time: 1.17 s
```

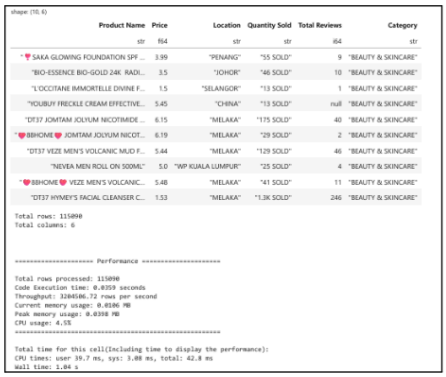
Figure 21: Output Screenshot of Dataset Integration by using Polars

shape: (10, 6)

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	str	str	str
"❣️SAKA GLOWING FOUNDATION SPF ...	3.99	"PENANG"	"55 SOLD"	"(9)"	"BEAUTY & SKINCARE"
"BIO-ESSENCE BIO-GOLD 24K RADI...	3.5	"JOHOR"	"46 SOLD"	"(10)"	"BEAUTY & SKINCARE"
"L'OCCITANE IMMORTELLE DIVINE F...	1.5	"SELANGOR"	"13 SOLD"	"(1)"	"BEAUTY & SKINCARE"
"YOUBUY FRECKLE CREAM EFFECTIVE...	5.45	"CHINA"	"13 SOLD"	null	"BEAUTY & SKINCARE"
"DT37 JOMTAM JOLYUM NICOTIMIDE ...	6.15	"MELAKA"	"175 SOLD"	"(40)"	"BEAUTY & SKINCARE"
"❣️88HOME❣️ JOMTAM JOLYUM NICOT...	6.19	"MELAKA"	"29 SOLD"	"(2)"	"BEAUTY & SKINCARE"
"DT37 VEZE MEN'S VOLCANIC MUD F...	5.44	"MELAKA"	"129 SOLD"	"(46)"	"BEAUTY & SKINCARE"
"NEVEA MEN ROLL ON 500ML"	5.0	"WP KUALA LUMPUR"	"25 SOLD"	"(4)"	"BEAUTY & SKINCARE"
"❣️88HOME❣️ VEZE MEN'S VOLCANIC...	5.48	"MELAKA"	"41 SOLD"	"(11)"	"BEAUTY & SKINCARE"
"DT37 HYMEY'S FACIAL CLEANSER C...	1.53	"MELAKA"	"1.3K SOLD"	"(246)"	"BEAUTY & SKINCARE"

```
Total rows: 115090
Total columns: 6


==================== Performance ====================

Total rows processed: 115090
Code Execution time: 0.0763 seconds
Throughput: 1507622.39 rows per second
Current memory usage: 0.0114 MB
Peak memory usage: 0.0413 MB
CPU usage: 4.0%
============================================================

Total time for this cell(Including time to display the performance):
CPU times: user 91.7 ms, sys: 21.8 ms, total: 113 ms
Wall time: 1.08 s
```

Figure 22: Output Screenshot of Standardization of String Data by using Polars

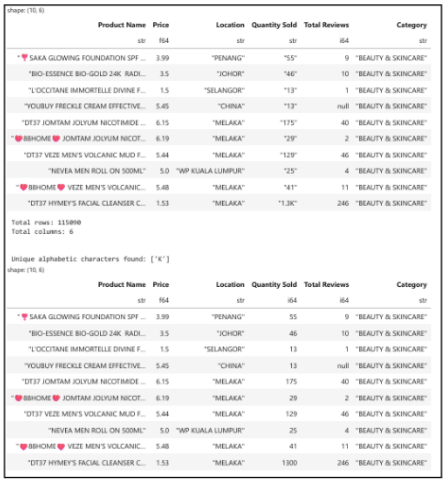


shape: (10, 6)

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	str	i64	str
* 💎 SAKA GLOWING FOUNDATION SPF ...	3.99	"PENANG"	"55 SOLD"	9	"BEAUTY & SKINCARE"
"BIO-ESSENCE BIO-GOLD 24K RADI...	3.5	"JOHOR"	"46 SOLD"	10	"BEAUTY & SKINCARE"
"L'OCCITANE IMMORTELLE DIVINE F...	1.5	"SELANGOR"	"13 SOLD"	1	"BEAUTY & SKINCARE"
"YOUBUY FRECKLE CREAM EFFECTIVE...	5.45	"CHINA"	"13 SOLD"	null	"BEAUTY & SKINCARE"
"DT37 JOMTAM JOLYUM NICOTIMIDE ...	6.15	"MELAKA"	"175 SOLD"	40	"BEAUTY & SKINCARE"
* ❤️BBHOME ❤️ JOMTAM JOLYUM NICOT...	6.19	"MELAKA"	"29 SOLD"	2	"BEAUTY & SKINCARE"
"DT37 VEZE MEN'S VOLCANIC MUD F...	5.44	"MELAKA"	"129 SOLD"	46	"BEAUTY & SKINCARE"
"NEVEA MEN ROLL ON 500ML"	5.0	"WP KUALA LUMPUR"	"25 SOLD"	4	"BEAUTY & SKINCARE"
* ❤️BBHOME ❤️ VEZE MEN'S VOLCANIC...	5.48	"MELAKA"	"41 SOLD"	11	"BEAUTY & SKINCARE"
"DT37 HYMEY'S FACIAL CLEANSER C...	1.53	"MELAKA"	"1.3K SOLD"	246	"BEAUTY & SKINCARE"

```
Total rows: 115090
Total columns: 6


==================== Performance ====================

Total rows processed: 115090
Code Execution time: 0.0359 seconds
Throughput: 3204586.72 rows per second
Current memory usage: 0.0106 MB
Peak memory usage: 0.0398 MB
CPU usage: 4.5%
============================================================

Total time for this cell(Including time to display the performance):
CPU times: user 39.7 ms, sys: 3.08 ms, total: 42.8 ms
Wall time: 1.04 s
```

Figure 23: Output Screenshot of Converting "Total Reviews" to int Data Type by using Polars



shape: (10, 6)

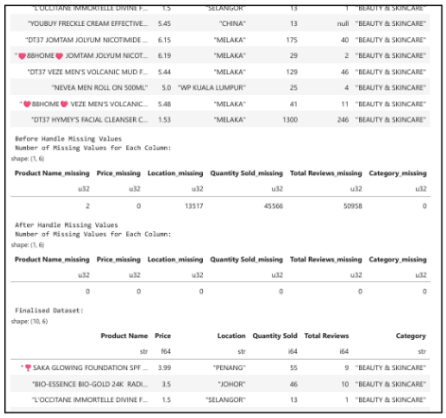
Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	str	i64	str
* 💎 SAKA GLOWING FOUNDATION SPF ...	3.99	"PENANG"	"55"	9	"BEAUTY & SKINCARE"
"BIO-ESSENCE BIO-GOLD 24K RADI...	3.5	"JOHOR"	"46"	10	"BEAUTY & SKINCARE"
"L'OCCITANE IMMORTELLE DIVINE F...	1.5	"SELANGOR"	"13"	1	"BEAUTY & SKINCARE"
"YOUBUY FRECKLE CREAM EFFECTIVE...	5.45	"CHINA"	"13"	null	"BEAUTY & SKINCARE"
"DT37 JOMTAM JOLYUM NICOTIMIDE ...	6.15	"MELAKA"	"175"	40	"BEAUTY & SKINCARE"
* ❤️BBHOME ❤️ JOMTAM JOLYUM NICOT...	6.19	"MELAKA"	"29"	2	"BEAUTY & SKINCARE"
"DT37 VEZE MEN'S VOLCANIC MUD F...	5.44	"MELAKA"	"129"	46	"BEAUTY & SKINCARE"
"NEVEA MEN ROLL ON 500ML"	5.0	"WP KUALA LUMPUR"	"25"	4	"BEAUTY & SKINCARE"
* ❤️BBHOME ❤️ VEZE MEN'S VOLCANIC...	5.48	"MELAKA"	"41"	11	"BEAUTY & SKINCARE"
"DT37 HYMEY'S FACIAL CLEANSER C...	1.53	"MELAKA"	"1.3K"	246	"BEAUTY & SKINCARE"

```
Total rows: 115090
Total columns: 6


Unique alphabetic characters found: ['K']
```

shape: (10, 6)

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	i64	i64	str
* 💎 SAKA GLOWING FOUNDATION SPF ...	3.99	"PENANG"	55	9	"BEAUTY & SKINCARE"
"BIO-ESSENCE BIO-GOLD 24K RADI...	3.5	"JOHOR"	46	10	"BEAUTY & SKINCARE"
"L'OCCITANE IMMORTELLE DIVINE F...	1.5	"SELANGOR"	13	1	"BEAUTY & SKINCARE"
"YOUBUY FRECKLE CREAM EFFECTIVE...	5.45	"CHINA"	13	null	"BEAUTY & SKINCARE"
"DT37 JOMTAM JOLYUM NICOTIMIDE ...	6.15	"MELAKA"	175	40	"BEAUTY & SKINCARE"
* ❤️BBHOME ❤️ JOMTAM JOLYUM NICOT...	6.19	"MELAKA"	29	2	"BEAUTY & SKINCARE"
"DT37 VEZE MEN'S VOLCANIC MUD F...	5.44	"MELAKA"	129	46	"BEAUTY & SKINCARE"
"NEVEA MEN ROLL ON 500ML"	5.0	"WP KUALA LUMPUR"	25	4	"BEAUTY & SKINCARE"
* ❤️BBHOME ❤️ VEZE MEN'S VOLCANIC...	5.48	"MELAKA"	41	11	"BEAUTY & SKINCARE"
"DT37 HYMEY'S FACIAL CLEANSER C...	1.53	"MELAKA"	1300	246	"BEAUTY & SKINCARE"

Figure 24: Output Screenshot of Converting "Quantity Sold" to int Data Type by using Polars



"L'OCCITANE IMMORTELLE DIVINE F...	1.5	"SELANGOR"	13	1	"BEAUTY & SKINCARE"
"YOUBUY FRECKLE CREAM EFFECTIVE...	5.45	"CHINA"	13	null	"BEAUTY & SKINCARE"
"DT37 JOMTAM JOLYUM NICOTIMIDE ...	6.15	"MELAKA"	175	40	"BEAUTY & SKINCARE"
* ❤️BBHOME ❤️ JOMTAM JOLYUM NICOT...	6.19	"MELAKA"	29	2	"BEAUTY & SKINCARE"
"DT37 VEZE MEN'S VOLCANIC MUD F...	5.44	"MELAKA"	129	46	"BEAUTY & SKINCARE"
"NEVEA MEN ROLL ON 500ML"	5.0	"WP KUALA LUMPUR"	25	4	"BEAUTY & SKINCARE"
* ❤️BBHOME ❤️ VEZE MEN'S VOLCANIC...	5.48	"MELAKA"	41	11	"BEAUTY & SKINCARE"
"DT37 HYMEY'S FACIAL CLEANSER C...	1.53	"MELAKA"	1300	246	"BEAUTY & SKINCARE"

```
Before Handle Missing Values
Number of Missing Values for Each Column:
```

shape: (1, 6)

Product Name_missing	Price_missing	Location_missing	Quantity Sold_missing	Total Reviews_missing	Category_missing
u32	u32	u32	u32	u32	u32
2	0	13517	45566	50958	0

```
After Handle Missing Values
Number of Missing Values for Each Column:
```

shape: (1, 6)

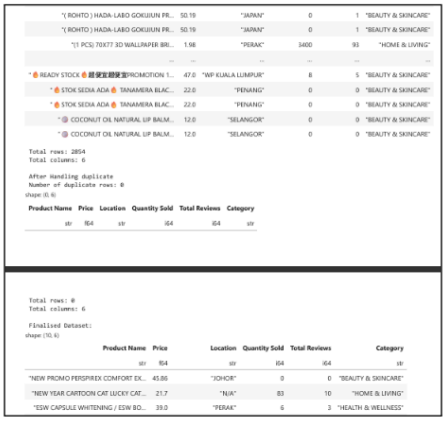
Product Name_missing	Price_missing	Location_missing	Quantity Sold_missing	Total Reviews_missing	Category_missing
u32	u32	u32	u32	u32	u32
0	0	0	0	0	0

```
Finalised Dataset:
```

shape: (10, 6)

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	i64	i64	str
* 💎 SAKA GLOWING FOUNDATION SPF ...	3.99	"PENANG"	55	9	"BEAUTY & SKINCARE"
"BIO-ESSENCE BIO-GOLD 24K RADI...	3.5	"JOHOR"	46	10	"BEAUTY & SKINCARE"
"L'OCCITANE IMMORTELLE DIVINE F...	1.5	"SELANGOR"	13	1	"BEAUTY & SKINCARE"

Figure 25: Output Screenshot of Checking and Handling Missing Values by using Polars



"(ROHTO) HADA-LABO GOKUJUN PR...	50.10	"JAPAN"	0	1	"BEAUTY & SKINCARE"
"(ROHTO) HADA-LABO GOKUJUN PR...	50.10	"JAPAN"	0	1	"BEAUTY & SKINCARE"
"[1 PCS] 70X77 3D WALLPAPER BRI...	1.98	"PERAK"	3400	93	"HOME & LIVING"
...	...	...	...	...	...
* 🔥 READY STOCK 🔥 超安全防安全PROMOTION 1...	47.0	"WP KUALA LUMPUR"	8	5	"BEAUTY & SKINCARE"
* 🔥 STOK SEDIA ADA 🔥 TANAMERA BLAC...	22.0	"PENANG"	0	0	"BEAUTY & SKINCARE"
* 🔥 STOK SEDIA ADA 🔥 TANAMERA BLAC...	22.0	"PENANG"	0	0	"BEAUTY & SKINCARE"
* 🥥 COCONUT OIL NATURAL LIP BALM...	12.0	"SELANGOR"	0	0	"BEAUTY & SKINCARE"
* 🥥 COCONUT OIL NATURAL LIP BALM...	12.0	"SELANGOR"	0	0	"BEAUTY & SKINCARE"

```
Total rows: 2854
Total columns: 6

After Handling duplicate
Number of duplicate rows: 0
```

shape: (0, 6)

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	i64	i64	str

```
Total rows: 0
Total columns: 6

Finalised Dataset:
```

shape: (10, 6)

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	i64	i64	str
"NEW PROMO PERSPIREX COMFORT EX...	45.86	"JOHOR"	0	0	"BEAUTY & SKINCARE"
"NEW YEAR CARTOON CAT LUCKY CAT...	21.7	"N/A"	83	10	"HOME & LIVING"
"ESW CAPSULE WHITENING / ESW BO...	39.0	"PERAK"	6	3	"HEALTH & WELLNESS"

Figure 26: Output Screenshot of Checking and Handling Duplicates by using Polars

```
====================  Performance  ====================

Total rows processed: 113596
Code Execution time: 0.0959 seconds
Throughput: 1184103.69 rows per second
Current memory usage: 0.0081 MB
Peak memory usage: 0.0143 MB
CPU usage: 4.0%
============================================================

Total time for this cell(Including time to display the performance):
CPU times: user 112 ms, sys: 33.7 ms, total: 145 ms
Wall time: 1.1 s
```

Figure 27: Output Screenshot of Exporting Cleaned Dataset File by using Polars

```
Min Price: 0.05
Max Price: 172.82

Group 1 (Budget Friendly Price): 65992 products
Group 2 (Affordable Price): 22626 products
Group 3 (Mid-Range Price): 16127 products
Group 4 (Premium Price): 14651 products

Group 1 (Budget Friendly Price):
```

shape: (10, 6)

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	i64	i64	str
"【MALAYSIA 3PIN PLUG】3L ELECTRI...	0.05	"SELANGOR"	0	0	"HOME APPLIANCES"
"LIVE TRACKING #NOT FOR SALES"	0.05	"KELANTAN"	0	1	"HOME & LIVING"
"SPOT BESEEN PLUS EYE CARE + BR...	0.1	"JOHOR"	0	0	"HEALTH & WELLNESS"
"1PCS 0.5MM BALL GEL INK PEN MA...	0.1	"SELANGOR"	33500	1354	"STATIONERY"
"HEALTH TREE REISSUE PRODUCT LI...	0.1	"CHINA"	10	3	"HEALTH & WELLNESS"
"BEFREE BESEEN PLUS VITAMIN EYE...	0.1	"JOHOR"	0	0	"HEALTH & WELLNESS"
"READY STOCK MICKEY MOUSE MASCO...	0.11	"WP KUALA LUMPUR"	49	1	"HOME & LIVING"
"✨ HARI RAYA HIASAN GANTUNG KRA...	0.12	"WP KUALA LUMPUR"	24	1	"HOME & LIVING"
"PENUTUP BOTTLE CAPS ARTEMIA,AR...	0.13	"PENANG"	7	0	"HEALTH & WELLNESS"
"《MILD EXFOLIATION》SOAP FOAM...	0.14	"PERAK"	922	20	"HEALTH & WELLNESS"

```
Total rows: 65992
Total columns: 6

📊 Category count:
- STATIONERY: 20460 products
- BEAUTY & SKINCARE: 17957 products
- HEALTH & WELLNESS: 10315 products
- HOME & LIVING: 8293 products
- WOMEN'S FASHION: 4653 products
- HOME APPLIANCES: 3837 products
- MOTHER & BABY: 477 products

Group 2 (Affordable Price):
```

Figure 28: Output Screenshot of Grouping Product into 4 Categories based on "Price" by using Polars

```
Minimum Number of Total Reviews: 0
Maximum Number of Total Reviews: 27

Group 1 (Least popular): 80729 products
Group 2 (Below Average Popularity): 7845 products
Group 3 (Above Average Popularity): 4653 products
Group 4 (Most popular): 2783 products

Group 1 (Least popular):
```

shape: (10, 6)

Product Name	Price	Location	Quantity Sold	Total Reviews	Category
str	f64	str	i64	i64	str
"UNICORN STATIONERY 0.5 MM RUB-...	2.0	"SELANGOR"	68	7	"STATIONERY"
"HOME LIVING ROOM BEDROOM FLOOR...	55.0	"NEGERI SEMBILAN"	28	7	"HOME & LIVING"
"ELBA 5.0L ELECTRIC KETTLE STAI...	119.0	"WP KUALA LUMPUR"	19	7	"HOME APPLIANCES"
"BURTS BEES 100% NATURAL MOISTU...	35.9	"PENANG"	44	7	"BEAUTY & SKINCARE"
"SOLID ADHESIVE NAIL GLUE SUPER...	2.86	"MELAKA"	31	7	"HEALTH & WELLNESS"
"🔥SHIP 24H🔥1PCS CURSIVE WRITING...	10.74	"CHINA"	16	7	"STATIONERY"
"QUICK EXTENDED GLUE MANICURE A...	4.56	"N/A"	81	7	"STATIONERY"
"** (笔芯0.5MM) 抄经金笔/ GOLD PEN / 5支...	3.5	"SELANGOR"	37	7	"STATIONERY"
"GLUMONY KAPSUL PEMUTIH BADAN W...	24.03	"WP KUALA LUMPUR"	20	7	"HEALTH & WELLNESS"

Figure 29: Output Screenshot of Grouping Products into 4 Categories based on "Total Reviews" by using Polars

shape: (10, 4)

Location	Total Quantity Sold	Average Price	Market Performance
str	i64	f64	f64
"SELANGOR"	10562173	100.849751	1.0652e9
"N/A"	18524452	43.647313	8.0854e8
"CHINA"	6341944	97.622898	6.1912e8
"WP KUALA LUMPUR"	1227922	101.519881	1.2466e8
"JOHOR"	1499826	78.073532	1.1710e8
"OVERSEAS"	2083568	48.480572	1.0101e8
"PENANG"	962481	100.382696	9.6616e7
"PERAK"	1029478	80.170026	8.2533e7
"HONG KONG"	376652	138.300619	5.2091e7
"KEDAH"	595993	80.050847	4.7710e7

```
====================  Performance  ====================

Total rows processed: 113596
Code Execution time: 0.1121 seconds
Throughput: 1013159.21 rows per second
Current memory usage: 0.0090 MB
Peak memory usage: 0.0222 MB
CPU usage: 100.0%
============================================================

Total time for this cell(Including time to display the performance):
CPU times: user 42 ms, sys: 15.1 ms, total: 57.1 ms
Wall time: 1.11 s
```

Figure 30: Output Screenshot of Evaluating and Ranking Market Performance based on "Quantity Sold" for each "Location" by using Polars

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	Upsee Women's Fashion Heat Resistant Long Curl...	17.13	China	753 sold	(188)	Women's Fashion
1	Summer women's fashionable high-end loose and ...	13.00	China	54 sold	(14)	Women's Fashion
2	Summer New Sports Suit Women's Fashion Slim Pr...	22.75	China	286 sold	(62)	Women's Fashion
3	Daring Backless V-Neck Dress Short Mini Skirt ...	19.72	None	270 sold	(37)	Women's Fashion
4	Chic Lace Birthday Dress Christmas Loose Sensa...	12.18	None	486 sold	(35)	Women's Fashion
5	LOMOGI Women's Fashion Dress + Outer Cardigan ...	15.23	China	1.6K sold	(455)	Women's Fashion
6	Summer 2024 Women's Knitted Long Hollow out Ve...	22.41	None	322 sold	(26)	Women's Fashion
7	Hot Princess Dress Set Sexy Slimming Waist Bod...	17.39	None	92 sold	(6)	Women's Fashion
8	LOMOGI Women's Fashion Dress + Outer CardigSle...	14.17	None	122 sold	(42)	Women's Fashion
9	Adult Latin Dance Skirt Square Dance Costume H...	20.55	None	350 sold	(58)	Women's Fashion

```
Total rows: 9698
Total columns: 6

---------- Performance ----------

Total rows processed: 115090
Code Execution time: 113.0075 seconds
Throughput: 1018.43 rows per second
Current memory usage: 10.9480 MB
Peak memory usage: 34.7047 MB
CPU usage: 84.8%

Total time for this cell (Including time to display the performance):
CPU times: user 1min 31s, sys: 469 ms, total: 1min 32s
Wall time: 1min 54s
```

Figure 31: Output Screenshot of Dataset Loading and Display by using PySpark

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	Penang	55 sold	(9)	Beauty & Skincare
1	Bio-Essence Bio-Gold 24k Radiance/Whitening/B...	3.50	Johor	46 sold	(10)	Beauty & Skincare
2	L'occitane Immortelle Divine Foaming Cleansing...	1.50	Selangor	13 sold	(1)	Beauty & Skincare
3	YOUBUY Freckle Cream Effectively Remove Melasm...	5.45	China	13 sold	None	Beauty & Skincare
4	DT37 Jomtam Jolyum Nicotimide Amino Acid Facia...	6.15	Melaka	175 sold	(40)	Beauty & Skincare
5	88Home VEZE Men's Volcanic Mud Facial Clea...	6.19	Melaka	29 sold	(2)	Beauty & Skincare
6	DT37 VEZE Men's Volcanic Mud Facial Cleanser B...	5.44	Melaka	129 sold	(46)	Beauty & Skincare
7	NEVEA MEN ROLL ON 500ML	5.00	Wp Kuala Lumpur	25 sold	(4)	Beauty & Skincare
8	88Home VEZE Men's Volcanic Mud Facial Clea...	5.48	Melaka	41 sold	(11)	Beauty & Skincare
9	DT37 HYMEY'S Facial Cleanser Cream Whitening M...	1.53	Melaka	1.3K sold	(246)	Beauty & Skincare

```
Total rows: 115090
Total columns: 6

---------- Performance ----------

Total rows processed: 115,090
Code Execution time: 4.8765 seconds
Throughput: 23,601.16 rows per second
Current memory usage: 0.0738 MB
Peak memory usage: 0.1984 MB
CPU usage: 12.6%

Total time for this cell (Including time to display the performance):
CPU times: user 176 ms, sys: 7 ms, total: 183 ms
Wall time: 5.88 s
```

Figure 32: Output Screenshot of Dataset Integration by using PySpark

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	PENANG	55 SOLD	(9)	BEAUTY & SKINCARE
1	BIO-ESSENCE BIO-GOLD 24K RADIANCE/WHITENING/B...	3.50	JOHOR	46 SOLD	(10)	BEAUTY & SKINCARE
2	L'OCCITANE IMMORTELLE DIVINE FOAMING CLEANSING...	1.50	SELANGOR	13 SOLD	(1)	BEAUTY & SKINCARE
3	YOUBUY FRECKLE CREAM EFFECTIVELY REMOVE MELASM...	5.45	CHINA	13 SOLD	None	BEAUTY & SKINCARE
4	DT37 JOMTAM JOLYUM NICOTIMIDE AMINO ACID FACIA...	6.15	MELAKA	175 SOLD	(40)	BEAUTY & SKINCARE
5	88HOME JOMTAM JOLYUM NICOTIMIDE AMINO ACID...	6.19	MELAKA	29 SOLD	(2)	BEAUTY & SKINCARE
6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129 SOLD	(46)	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 500ML	5.00	WP KUALA LUMPUR	25 SOLD	(4)	BEAUTY & SKINCARE
8	88HOME VEZE MEN'S VOLCANIC MUD FACIAL CLEA...	5.48	MELAKA	41 SOLD	(11)	BEAUTY & SKINCARE
9	DT37 HYMEY'S FACIAL CLEANSER CREAM WHITENING M...	1.53	MELAKA	1.3K SOLD	(246)	BEAUTY & SKINCARE

```
Total rows: 115090
Total columns: 6

---------- Performance ----------

Total rows processed: 115,090
Code Execution time: 4.2217 seconds
Throughput: 27,268.92 rows per second
Current memory usage: 0.0679 MB
Peak memory usage: 0.1824 MB
CPU usage: 61.6%

Total time for this cell (Including time to display the performance):
CPU times: user 205 ms, sys: 12 ms, total: 217 ms
Wall time: 5.23 s
```

Figure 33: Output Screenshot of Standardization of String Data by using PySpark

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	PENANG	55 SOLD	9.0	BEAUTY & SKINCARE
1	BIO-ESSENCE BIO-GOLD 24K RADIANCE/WHITENING/B...	3.50	JOHOR	46 SOLD	10.0	BEAUTY & SKINCARE
2	L'OCCITANE IMMORTELLE DIVINE FOAMING CLEANSING...	1.50	SELANGOR	13 SOLD	1.0	BEAUTY & SKINCARE
3	YOUBUY FRECKLE CREAM EFFECTIVELY REMOVE MELASM...	5.45	CHINA	13 SOLD	NaN	BEAUTY & SKINCARE
4	DT37 JOMTAM JOLYUM NICOTIMIDE AMINO ACID FACIA...	6.15	MELAKA	175 SOLD	40.0	BEAUTY & SKINCARE
5	88HOME JOMTAM JOLYUM NICOTIMIDE AMINO ACID...	6.19	MELAKA	29 SOLD	2.0	BEAUTY & SKINCARE
6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129 SOLD	46.0	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 500ML	5.00	WP KUALA LUMPUR	25 SOLD	4.0	BEAUTY & SKINCARE
8	88HOME VEZE MEN'S VOLCANIC MUD FACIAL CLEA...	5.48	MELAKA	41 SOLD	11.0	BEAUTY & SKINCARE
9	DT37 HYMEY'S FACIAL CLEANSER CREAM WHITENING M...	1.53	MELAKA	1.3K SOLD	246.0	BEAUTY & SKINCARE

```
Total rows: 115090
Total columns: 6

---------- Performance ----------

Total rows processed: 115,090
Code Execution time: 5.1355 seconds
Throughput: 22,410.57 rows per second
Current memory usage: 0.0748 MB
Peak memory usage: 0.1679 MB
CPU usage: 21.6%

Total time for this cell (Including time to display the performance):
CPU times: user 210 ms, sys: 11.9 ms, total: 222 ms
Wall time: 6.14 s
```

Figure 34: Output Screenshot of Converting "Total Reviews" to int Data Type by using PySpark

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	PENANG	55	9.0	BEAUTY & SKINCARE
1	BIO-ESSENCE BIO-GOLD 24K RADIANCE/WHITENING/B...	3.50	JOHOR	46	10.0	BEAUTY & SKINCARE
2	L'OCCITANE IMMORTELLE DIVINE FOAMING CLEANSING...	1.50	SELANGOR	13	1.0	BEAUTY & SKINCARE
3	YOUBUY FRECKLE CREAM EFFECTIVELY REMOVE MELASM...	5.45	CHINA	13	NaN	BEAUTY & SKINCARE
4	DT37 JOMTAM JOLYUM NICOTIMIDE AMINO ACID FACIA...	6.15	MELAKA	175	40.0	BEAUTY & SKINCARE
5	88HOME JOMTAM JOLYUM NICOTIMIDE AMINO ACID...	6.19	MELAKA	29	2.0	BEAUTY & SKINCARE
6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129	46.0	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 500ML	5.00	WP KUALA LUMPUR	25	4.0	BEAUTY & SKINCARE
8	88HOME VEZE MEN'S VOLCANIC MUD FACIAL CLEA...	5.48	MELAKA	41	11.0	BEAUTY & SKINCARE
9	DT37 HYMEY'S FACIAL CLEANSER CREAM WHITENING M...	1.53	MELAKA	1300	246.0	BEAUTY & SKINCARE

```
Total rows: 115090
Total columns: 6

---------- Performance ----------

Total rows processed: 115,090
Code Execution time: 4.1328 seconds
Throughput: 27,847.80 rows per second
Current memory usage: 0.0972 MB
Peak memory usage: 0.1938 MB
CPU usage: 6.8%

Total time for this cell (Including time to display the performance):
CPU times: user 213 ms, sys: 8.93 ms, total: 222 ms
Wall time: 5.14 s
```

Figure 35: Output Screenshot of Converting "Quantity Sold" to int Data Type by using PySpark

6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129	46.0	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 500ML	5.00	WP KUALA LUMPUR	25	4.0	BEAUTY & SKINCARE
8	88HOME VEZE MEN'S VOLCANIC MUD FACIAL CLEA...	5.48	MELAKA	41	11.0	BEAUTY & SKINCARE
9	DT37 HYMEY'S FACIAL CLEANSER CREAM WHITENING M...	1.53	MELAKA	1300	246.0	BEAUTY & SKINCARE

```
Before Handling Missing Values
Missing Values:
```

	Product Name_missing	Price_missing	Location_missing	Quantity Sold_missing	Total Reviews_missing	Category_missing
0	2	0	13517	45566	50958	0

```
After Handling Missing Values
Missing Values:
```

	Product Name_missing	Price_missing	Location_missing	Quantity Sold_missing	Total Reviews_missing	Category_missing
0	0	0	0	0	0	0

```
Finalised Dataset:
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	SAKA GLOWING FOUNDATION SPF 50+ / FW FATIN W...	3.99	PENANG	55	9	BEAUTY & SKINCARE
1	BIO-ESSENCE BIO-GOLD 24K RADIANCE/WHITENING/B...	3.50	JOHOR	46	10	BEAUTY & SKINCARE
2	L'OCCITANE IMMORTELLE DIVINE FOAMING CLEANSING...	1.50	SELANGOR	13	1	BEAUTY & SKINCARE
3	YOUBUY FRECKLE CREAM EFFECTIVELY REMOVE MELASM...	5.45	CHINA	13	0	BEAUTY & SKINCARE
4	DT37 JOMTAM JOLYUM NICOTIMIDE AMINO ACID FACIA...	6.15	MELAKA	175	40	BEAUTY & SKINCARE
5	88HOME JOMTAM JOLYUM NICOTIMIDE AMINO ACID...	6.19	MELAKA	29	2	BEAUTY & SKINCARE
6	DT37 VEZE MEN'S VOLCANIC MUD FACIAL CLEANSER B...	5.44	MELAKA	129	46	BEAUTY & SKINCARE
7	NEVEA MEN ROLL ON 500ML	5.00	WP KUALA LUMPUR	25	4	BEAUTY & SKINCARE

Figure 36: Output Screenshot of Checking and Handling Missing Values by using PySpark

3	(ROHTO) HADA-LABO GOKUJUN PREMIUM HYALURONIC...	50.19	JAPAN	0	1	BEAUTY & SKINCARE
4	(1 PCS) 70X77 3D WALLPAPER BRICK WALL STICKERS...	1.98	PERAK	3400	93	HOME & LIVING
...	...	...	...	...	...	...
2849	READY STOCK 超级宝超便宜PROMOTION 100% ORIGINAL 美乐家...	47.00	WP KUALA LUMPUR	8	5	BEAUTY & SKINCARE
2850	STOK SEDIA ADA TANAMERA BLACK FORMULATION FA...	22.00	PENANG	0	0	BEAUTY & SKINCARE
2851	STOK SEDIA ADA TANAMERA BLACK FORMULATION FA...	22.00	PENANG	0	0	BEAUTY & SKINCARE
2852	COCONUT OIL NATURAL LIP BALM สินค้าพร้อมส่ง!...	12.00	SELANGOR	0	0	BEAUTY & SKINCARE
2853	COCONUT OIL NATURAL LIP BALM สินค้าพร้อมส่ง!...	12.00	SELANGOR	0	0	BEAUTY & SKINCARE

2854 rows × 6 columns

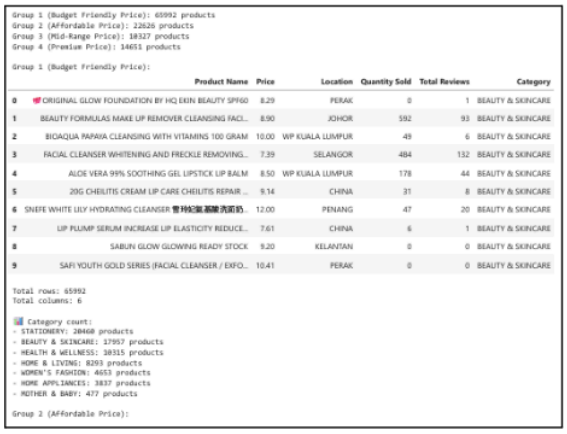
```
Total rows: 2854
Total columns: 6

After Handling Duplicates
Number of duplicate rows: 0
```

Product Name	Price	Location	Quantity Sold	Total Reviews	Category

```
Total rows: 0
Total columns: 6

Finalised Dataset:
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	ORIGINAL GLOW FOUNDATION BY HQ EKIN BEAUTY SPF60	8.29	PERAK	0	1	BEAUTY & SKINCARE
1	BEAUTY FORMULAS MAKE UP REMOVER CLEANSING FACI...	8.90	JOHOR	592	93	BEAUTY & SKINCARE
2	BIOAQUA PAPAYA CLEANSING WITH VITAMINS 100 GRAM	10.00	WP KUALA LUMPUR	49	6	BEAUTY & SKINCARE
3	FACIAL CLEANSER WHITENING AND FRECKLE REMOVING...	7.39	SELANGOR	484	132	BEAUTY & SKINCARE
4	ALOE VERA 99% SOOTHING GEL LIPSTICK LIP BALM	8.50	WP KUALA LUMPUR	178	44	BEAUTY & SKINCARE
5	20G CHEILITIS CREAM LIP CARE CHEILITIS REPAIR ...	9.14	CHINA	31	8	BEAUTY & SKINCARE
6	SNEFE WHITE LILY HYDRATING CLEANSER 雪玲妃氨基酸洗面奶...	12.00	PENANG	47	20	BEAUTY & SKINCARE

Figure 37: Output Screenshot of Checking and Handling Duplicates by using PySpark



```
Group 1 (Budget Friendly Price): 65992 products
Group 2 (Affordable Price): 22626 products
Group 3 (Mid-Range Price): 10327 products
Group 4 (Premium Price): 14651 products

Group 1 (Budget Friendly Price):
```

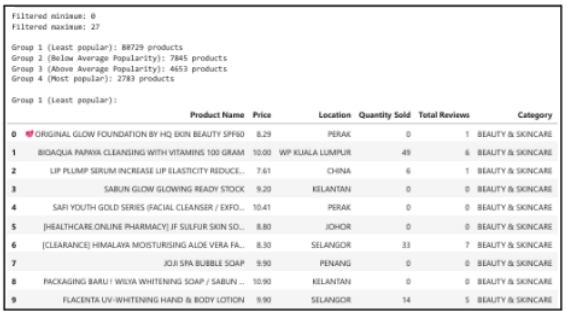
	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	ORIGINAL GLOW FOUNDATION BY HQ EKIN BEAUTY SPF60	8.29	PERAK	0	1	BEAUTY & SKINCARE
1	BEAUTY FORMULAS MAKE UP REMOVER CLEANSING FACI...	8.90	JOHOR	592	93	BEAUTY & SKINCARE
2	BIOAQUA PAPAYA CLEANSING WITH VITAMINS 100 GRAM	10.00	WP KUALA LUMPUR	49	6	BEAUTY & SKINCARE
3	FACIAL CLEANSER WHITENING AND FRECKLE REMOVING...	7.39	SELANGOR	484	132	BEAUTY & SKINCARE
4	ALOE VERA 99% SOOTHING GEL LIPSTICK LIP BALM	8.50	WP KUALA LUMPUR	178	44	BEAUTY & SKINCARE
5	20G CHEILITIS CREAM LIP CARE CHEILITIS REPAIR ...	9.14	CHINA	31	8	BEAUTY & SKINCARE
6	SNEFE WHITE LILY HYDRATING CLEANSER 雪玲妃氨基酸洗面奶...	12.00	PENANG	47	20	BEAUTY & SKINCARE
7	LIP PLUMP SERUM INCREASE LIP ELASTICITY REDUCE...	7.61	CHINA	6	1	BEAUTY & SKINCARE
8	SABUN GLOW GLOWING READY STOCK	9.20	KELANTAN	0	0	BEAUTY & SKINCARE
9	SAFI YOUTH GOLD SERIES (FACIAL CLEANSER / EXFO...	10.41	PERAK	0	0	BEAUTY & SKINCARE

```
Total rows: 65992
Total columns: 6

Category count:
- STATIONERY: 20460 products
- BEAUTY & SKINCARE: 17957 products
- HEALTH & WELLNESS: 10315 products
- HOME & LIVING: 8293 products
- WOMEN'S FASHION: 4653 products
- HOME APPLIANCES: 3837 products
- MOTHER & BABY: 477 products

Group 2 (Affordable Price):
```

Figure 39: Output Screenshot of Grouping Product into 4 Categories based on "Price" by using PySpark

```
====================== Performance ======================

Total rows processed: 113596
Code Execution time: 10.6576 seconds
Throughput: 10658.73 rows per second
Current memory usage: 0.0334 MB
Peak memory usage: 0.0880 MB
CPU usage: 75.1%
=========================================================

Total time for this cell(Including time to display the performance):
CPU times: user 140 ms, sys: 8.15 ms, total: 148 ms
Wall time: 11.7 s
```

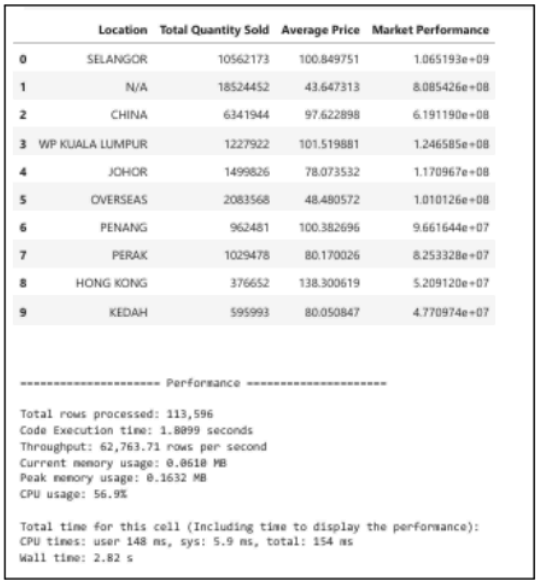
Figure 38: Output Screenshot of Exporting Cleaned Dataset File by using PySpark



```
Filtered minimum: 0
Filtered maximum: 27

Group 1 (Least popular): 88729 products
Group 2 (Below Average Popularity): 7845 products
Group 3 (Above Average Popularity): 4653 products
Group 4 (Most popular): 2783 products

Group 1 (least popular):
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	ORIGINAL GLOW FOUNDATION BY HQ EKIN BEAUTY SPF60	8.29	PERAK	0	1	BEAUTY & SKINCARE
1	BIOAQUA PAPAYA CLEANSING WITH VITAMINS 100 GRAM	10.00	WP KUALA LUMPUR	49	6	BEAUTY & SKINCARE
2	LIP PLUMP SERUM INCREASE LIP ELASTICITY REDUCE...	7.61	CHINA	6	1	BEAUTY & SKINCARE
3	SABUN GLOW GLOWING READY STOCK	9.20	KELANTAN	0	0	BEAUTY & SKINCARE
4	SAFI YOUTH GOLD SERIES (FACIAL CLEANSER / EXFO...	10.41	PERAK	0	0	BEAUTY & SKINCARE
5	[HEALTHCARE ONLINE PHARMACY] JF SULFUR SKIN SO...	8.80	JOHOR	0	0	BEAUTY & SKINCARE
6	[CLEARANCE] HIMALAYA MOISTURISING ALOE VERA FA...	8.30	SELANGOR	33	7	BEAUTY & SKINCARE
7	JOJI SPA BUBBLE SOAP	9.90	PENANG	0	0	BEAUTY & SKINCARE
8	PACKAGING BARU ! WILYA WHITENING SOAP / SABUN ...	10.90	KELANTAN	0	0	BEAUTY & SKINCARE
9	PLACENTA UV-WHITENING HAND & BODY LOTION	9.90	SELANGOR	14	5	BEAUTY & SKINCARE

Figure 40: Output Screenshot of Grouping Products into 4 Categories based on "Total Reviews" by using PySpark



	Location	Total Quantity Sold	Average Price	Market Performance
0	SELANGOR	10562173	100.849751	1.065193e+09
1	N/A	18524452	43.647313	8.085426e+08
2	CHINA	6341944	97.622898	6.191190e+08
3	WP KUALA LUMPUR	1227922	101.519881	1.246585e+08
4	JOHOR	1499826	78.073532	1.170967e+08
5	OVERSEAS	2083568	48.480572	1.010126e+08
6	PENANG	962481	100.382696	9.661644e+07
7	PERAK	1029478	80.170026	8.253328e+07
8	HONG KONG	376652	138.300619	5.209120e+07
9	KEDAH	595993	80.050847	4.770974e+07

```
==================== Performance ====================

Total rows processed: 113,596
Code Execution time: 1.8099 seconds
Throughput: 62,763.71 rows per second
Current memory usage: 0.0610 MB
Peak memory usage: 0.1632 MB
CPU usage: 56.9%

Total time for this cell (Including time to display the performance):
CPU times: user 148 ms, sys: 5.9 ms, total: 154 ms
Wall time: 2.82 s
```

Figure 41: Output Screenshot of Evaluating and Ranking Market Performance based on "Quantity Sold" for each "Location" by using PySpark

Links to full code repo or dataset

1. **Pandas:**
   - Part 1
   - Part 2
2. **Polars:**
   - Part 1
   - Part 2
3. **PySpark:**
   - Part 1
   - Part 2
4. **Dataset**
   - Raw Dataset
   - Cleaned Dataset

HPDP_Project Report.pdf

ORIGINALITY REPORT

3%	2%	0%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	neptune.ai Internet Source	1%
2	www.diva-portal.org Internet Source	<1%
3	2captcha.com Internet Source	<1%
4	Submitted to University of Surrey Student Paper	<1%
5	Submitted to Universiti Teknologi Malaysia Student Paper	<1%
6	documents1.worldbank.org Internet Source	<1%
7	Submitted to Arab Open University Student Paper	<1%
8	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelligent Computing and Communication Techniques - Volume 2", CRC Press, 2025 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		