

Part 2 Data Optimization

Prepared by:

Step 1 : Install and Import Libraries

```
In [ ]: 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

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

Step 2 : Upload Excel Files

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

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving pyspark\_cleaned\_dataset.csv to pyspark\_cleaned\_dataset.csv

Step 3 : Load and Display Dataset, Checking on Total Numbers of Rows and Columns

```
In [ ]: %%time

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

filename = list(uploaded.keys())[0]

df_cleaned = spark.read.csv(
    filename,
    header=True,
    inferSchema=True,
    quote='',
    escape='',
    multiline=True
)

display(df_cleaned.limit(10).toPandas())

total_rows = df_cleaned.count()

total_columns = len(df_cleaned.columns)

print(f"Total rows: {total_rows}")
print(f"Total columns: {total_columns}")

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):")
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	<del>100</del> 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: 113596  
Total columns: 6  
===== Performance =====  
  
Total rows processed: 113596  
Code Execution time: 13.9451 seconds  
Throughput: 8145.97 rows per second  
Current memory usage: 3.5778 MB  
Peak memory usage: 3.9638 MB  
CPU usage: 29.5%  
=====

Total time for this cell(Including time to display the performance):  
CPU times: user 824 ms, sys: 43.3 ms, total: 867 ms  
Wall time: 15 s

Step 4 : Dividing Products into 4 Categories Based on Price using Pandas

```
In [ ]: %%time

tracemalloc.start()
start_time = time.perf_counter()
total_rows = df_cleaned.count()

df_price = df_cleaned.filter(F.col("Price") > 0)

q1 = df_price.approxQuantile("Price", [0.25], 0.0)[0]
q3 = df_price.approxQuantile("Price", [0.75], 0.0)[0]
iqr = q3 - q1
lower_bound = q1 - 1.5 * iqr
upper_bound = q3 + 1.5 * iqr

df_cleared = df_price.filter((F.col("Price") >= lower_bound) & (F.col("Price") <= upper_bound))

min_price = df_cleared.agg(F.min("Price")).collect()[0][0]
max_price = df_cleared.agg(F.max("Price")).collect()[0][0]

print(f"Min Price: {min_price}")
print(f"Max Price: {max_price}")

price_range = (max_price - min_price) / 4
bound1 = min_price + price_range
bound2 = min_price + 2 * price_range
bound3 = min_price + 3 * price_range

group1 = df_price.filter(F.col("Price") <= bound1)
group2 = df_price.filter((F.col("Price") > bound1) & (F.col("Price") <= bound2))
group3 = df_price.filter((F.col("Price") > bound2) & (F.col("Price") <= bound3))
group4 = df_price.filter(F.col("Price") > bound3)

print(f"\nGroup 1 (Budget Friendly Price): {group1.count()} products")
print(f"Group 2 (Affordable Price): {group2.count()} products")
print(f"Group 3 (Mid-Range Price): {group3.count()} products")
print(f"Group 4 (Premium Price): {group4.count()} products")

def print_category_counts(df):
    print("\nCategory count:")
    category_counts = df.groupBy("Category").count().orderBy("count", ascending=False)
    for row in category_counts.collect():
        print(f"- {row['Category']}: {row['count']} products")

print("\nGroup 1 (Budget Friendly Price):")
display(group1.limit(10).toPandas())
print(f"Total rows: {group1.count()}")
print(f"Total columns: {len(group1.columns)}\n")
print_category_counts(group1)
```

```
print("\nGroup 2 (Affordable Price):")
display(group2.limit(10).toPandas())
print(f"Total rows: {group2.count()}")
print(f"Total columns: {len(group2.columns)}\n")
print_category_counts(group2)

print("\nGroup 3 (Mid-Range Price):")
display(group3.limit(10).toPandas())
print(f"Total rows: {group3.count()}")
print(f"Total columns: {len(group3.columns)}\n")
print_category_counts(group3)

print("\nGroup 4 (Premium Price):")
display(group4.limit(10).toPandas())
print(f"Total rows: {group4.count()}")
print(f"Total columns: {len(group4.columns)}\n")
print_category_counts(group4)

total_rows = df_price.count()

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

execution_time = end_time - start_time
throughput = total_rows / execution_time if execution_time > 0 else 0

print("\n===== 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):")
```


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): 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):

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	AR COLLAGEN PLUS DIETARY SUPPLEMENT PRODUCT	45.00	KELANTAN	7	2	BEAUTY & SKINCARE
1	[READY STOK] FRESKIN WHITE WHITENING BODY SHOW...	47.00	JOHOR	0	1	BEAUTY & SKINCARE
2	BODY BUTTER STICK MOISTURIZING BODY CREAM MOIS...	45.84	CHINA	0	0	BEAUTY & SKINCARE
3	CETAPHIL HYDRATING FOAMING CREAM CLEANSER 236ML	43.90	SELANGOR	0	0	BEAUTY & SKINCARE
4	PAPULEX MOUSSANT SOAP FREE CLEANSING GEL 150ML...	69.21	KEDAH	830	203	BEAUTY & SKINCARE
5	DERMAREL LIPID REPLENISHING CLEANSER 400ML   S...	50.90	SELANGOR	277	53	BEAUTY & SKINCARE
6	*RJ BEAUTY BOOSTER WHITENING SOAP	61.68	SELANGOR	0	0	BEAUTY & SKINCARE
7	EXP AUG 2025 DR.WU OFFICIAL RENEWAL CLEANSING ...	72.50	SELANGOR	160	29	BEAUTY & SKINCARE
8	TAIWAN AMINO ACID CLEANSING MILK TRUU 76 YEAST...	64.60	CHINA	0	0	BEAUTY & SKINCARE
9	THE FACE SHOP DAILY PERFUMED FOAM CLEANSER HOL...	49.00	NEGERI SEMBILAN	0	1	BEAUTY & SKINCARE


Total rows: 22626  
Total columns: 6

 Category count:  
- HEALTH & WELLNESS: 6380 products  
- BEAUTY & SKINCARE: 5373 products  
- WOMEN'S FASHION: 3623 products  
- HOME APPLIANCES: 3053 products  
- HOME & LIVING: 2309 products  
- STATIONERY: 1170 products  
- MOTHER & BABY: 718 products

Group 3 (Mid-Range Price):

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	APRIL22 SKIN ENERGY PURIFYING CLEANSER 120G能量活...	96.00	JOHOR	6	4	BEAUTY & SKINCARE
1	TEOXANE RHA™ MICELLAR SOLUTION 200ML	129.00	SELANGOR	14	0	BEAUTY & SKINCARE
2	NUXE REVE DE MIEL FACE CLEANSING & MAKEUP REMO...	96.10	CHINA	12	0	BEAUTY & SKINCARE
3	PAULA'S CHOICE SKIN BALANCING OIL-REDUCING CLE...	108.00	SELANGOR	136	43	BEAUTY & SKINCARE
4	CLINIQUE RINSE-OFF FOAMING CLEANSER 150ML	93.69	CHINA	14	4	BEAUTY & SKINCARE
5	LA ROCHE-POSAY TOLERIANE PURIFYING FOAMING CRE...	103.22	SOUTH KOREA	0	0	BEAUTY & SKINCARE
6	SHISEIDO MEN CLEANSING FOAM 125ML	92.40	HONG KONG	148	41	BEAUTY & SKINCARE
7	FRESH - SUGAR LIP TREATMENT - COCOA 4.3G/0.15OZ	107.13	CHINA	0	0	BEAUTY & SKINCARE
8	NU SKIN NUSKIN LUMISPA TREATMENT HEAD - READY ...	129.00	SELANGOR	0	0	BEAUTY & SKINCARE
9	PURA D'OR ORGANIC BEARD OIL, MUSTACHE CARE, MI...	87.82	SELANGOR	5	0	BEAUTY & SKINCARE


Total rows: 10327  
Total columns: 6

 Category count:  
- HEALTH & WELLNESS: 3866 products  
- BEAUTY & SKINCARE: 2008 products  
- HOME APPLIANCES: 1938 products  
- HOME & LIVING: 977 products  
- WOMEN'S FASHION: 774 products  
- STATIONERY: 451 products  
- MOTHER & BABY: 313 products

Group 4 (Premium Price):

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	LOTUS HERBALS RADIANT PEARL CELLULAR FACIAL KI...	130.00	SELANGOR	0	0	BEAUTY & SKINCARE
1	ESTEE LAUDER PERFECTLY CLEAN MULTI-ACTION FOAM...	165.09	JOHOR	0	0	BEAUTY & SKINCARE
2	D'ALBA   MILD SKIN BALANCING VEGAN CLEANSER (2...	148.03	SOUTH KOREA	0	0	BEAUTY & SKINCARE
3	PERSONALIZED MINI SKINCARE BEAUTY FRIDGE 8L - ...	160.00	PERAK	0	0	BEAUTY & SKINCARE
4	CAMELLA APRICOT FACIAL FOAMING CLEANSER	144.90	SELANGOR	126	20	BEAUTY & SKINCARE
5	BIOTHERM FORCE SUPREME ANTI-AGING CLEANSER 125ML	234.80	HONG KONG	40	12	BEAUTY & SKINCARE
6	[NSC] S.LITE 60 CAPSULES 450MG (SHORT EXPIRY N...	245.17	WP KUALA LUMPUR	0	0	BEAUTY & SKINCARE
7	LA PRAIRIE PURIFYING CREAM CLEANSER 200ML	268.30	CHINA	0	1	BEAUTY & SKINCARE
8	HANA SAPPHIRE AIR IPL HAIR REMOVAL MACHINE DEV...	967.13	SELANGOR	0	1	BEAUTY & SKINCARE
9	PEVONIA MALAYSIA - LUMAFIRM® BODY MOISTURIZER ...	380.00	SELANGOR	0	1	BEAUTY & SKINCARE

Total rows: 14651  
Total columns: 6

 Category count:  
- HOME APPLIANCES: 4700 products  
- HEALTH & WELLNESS: 3581 products  
- BEAUTY & SKINCARE: 2388 products  
- HOME & LIVING: 1642 products  
- MOTHER & BABY: 923 products  
- STATIONERY: 794 products  
- WOMEN'S FASHION: 623 products

===== Performance =====

Total rows processed: 113,596  
Code Execution time: 22.6163 seconds  
Throughput: 5,022.74 rows per second  
Current memory usage: 0.2022 MB  
Peak memory usage: 0.3516 MB  
CPU usage: 5.5%

Total time for this cell (Including time to display the performance):  
CPU times: user 1.03 s, sys: 43.9 ms, total: 1.07 s  
Wall time: 23.6 s

Step 5 : Filtering Based on 'Total Reviews' to Determine Popularity of Products

```
In [ ]: %%time

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

Q1 = df_cleaned.approxQuantile("Total Reviews", [0.25], 0.0)[0]
Q3 = df_cleaned.approxQuantile("Total Reviews", [0.75], 0.0)[0]
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

df_filtered = df_cleaned.filter(
    (F.col("Total Reviews") >= lower_bound) & (F.col("Total Reviews") <= upper_bound)
)


min_ratings = df_filtered.agg(F.min("Total Reviews")).collect()[0][0]
max_ratings = df_filtered.agg(F.max("Total Reviews")).collect()[0][0]

print(f"Filtered minimum: {min_ratings}")
print(f"Filtered maximum: {max_ratings}")

rating_range = round((max_ratings - min_ratings) / 4)
bound1 = min_ratings + rating_range
bound2 = min_ratings + 2 * rating_range
bound3 = min_ratings + 3 * rating_range

group1 = df_filtered.filter(F.col("Total Reviews") <= bound1)
group2 = df_filtered.filter((F.col("Total Reviews") > bound1) & (F.col("Total Reviews") <= bound2))
group3 = df_filtered.filter((F.col("Total Reviews") > bound2) & (F.col("Total Reviews") <= bound3))
group4 = df_filtered.filter(F.col("Total Reviews") > bound3)

print(f"\nGroup 1 (Least popular): {group1.count()} products")
print(f"Group 2 (Below Average Popularity): {group2.count()} products")
print(f"Group 3 (Above Average Popularity): {group3.count()} products")
print(f"Group 4 (Most popular): {group4.count()} products")

def print_category_counts(df):
    print("\n Category count:")
    category_counts = df.groupBy("Category").count().orderBy("count", ascending=False)
```



```
for row in category_counts.collect():
    print(f"- {row['Category']}: {row['count']} products")

print("\nGroup 1 (Least popular):")
display(group1.limit(10).toPandas())
print(f"Total rows: {group1.count()}")
print(f"Total columns: {len(group1.columns)}\n")
print_category_counts(group1)

print("\nGroup 2 (Below Average Popularity):")
display(group2.limit(10).toPandas())
print(f"Total rows: {group2.count()}")
print(f"Total columns: {len(group2.columns)}\n")
print_category_counts(group2)

print("\nGroup 3 (Above Average Popularity):")
display(group3.limit(10).toPandas())
print(f"Total rows: {group3.count()}")
print(f"Total columns: {len(group3.columns)}\n")
print_category_counts(group3)

print("\nGroup 4 (Most popular):")
display(group4.limit(10).toPandas())
print(f"Total rows: {group4.count()}")
print(f"Total columns: {len(group4.columns)}\n")
print_category_counts(group4)

total_rows = df_cleaned.count()

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

execution_time = end_time - start_time
throughput = total_rows / execution_time if execution_time > 0 else 0

print("\n===== 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):")
```

Filtered minimum: 0  
Filtered maximum: 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):

	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	FLACENTA UV-WHITENING HAND & BODY LOTION	9.90	SELANGOR	14	5	BEAUTY & SKINCARE

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
0	20G CHEILITIS CREAM LIP CARE CHEILITIS REPAIR ...	9.14	CHINA	31	8	BEAUTY & SKINCARE
1	MODELONES GEL NAIL POLISH SET POPULAR NUDE NEU...	19.72	CHINA	55	9	BEAUTY & SKINCARE
2	GMEELAN CENTELLA BHA SMOOTHING GEL CLEANSER CE...	30.00	SELANGOR	30	12	BEAUTY & SKINCARE
3	HADA LABO AGING CARE FACE WASH 100G	24.97	SELANGOR	33	13	BEAUTY & SKINCARE
4	1KG GOAT MILK SOAP BASE (CONTAIN GLYCERIN)   1...	29.90	WP KUALA LUMPUR	38	10	BEAUTY & SKINCARE
5	WHITENING BODY LOTION GLOW SKIN WHITE FOR FULL...	26.37	PERAK	26	10	BEAUTY & SKINCARE
6	HANBOLI PEPTIDE EYE ESSENCE STICK SPRING SUMME...	11.30	N/A	190	11	BEAUTY & SKINCARE
7	NIVEA WOMEN DEODORANT EXTRA BRIGHT C&E 50ML	9.90	PERAK	30	8	BEAUTY & SKINCARE
8	BEDAK PONDS PINKISH GLOW 100% ORIGINAL - BEST ...	9.99	PERLIS	74	11	BEAUTY & SKINCARE
9	ASTRAGALUS CREAM FACIAL CLEANSER TRANSPARENT M...	12.50	N/A	1800	8	BEAUTY & SKINCARE

Total rows: 7845  
Total columns: 6

- 🏷️ Category count:
- BEAUTY & SKINCARE: 2205 products
  - HEALTH & WELLNESS: 1784 products
  - STATIONERY: 1752 products
  - HOME APPLIANCES: 864 products
  - HOME & LIVING: 791 products
  - WOMEN'S FASHION: 408 products
  - MOTHER & BABY: 41 products

Group 3 (Above Average Popularity):

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	SNEFE WHITE LILY HYDRATING CLEANSER 雪玲妃氨基酸洗面奶...	12.00	PENANG	47	20	BEAUTY & SKINCARE
1	硫磺除螨祛痘洗护套装 QINGLING SULFUR FACIAL CLEANSER +BO...	15.50	JOHOR	417	16	BEAUTY & SKINCARE
2	EELHOE HAIR DYE GRAY COLOR FASHION GREY DYE HA...	16.50	CHINA	26	20	BEAUTY & SKINCARE
3	CLEAR MEN DEEP CLEANSE ANTI-DANDRUFF SHAMPOO 3...	18.04	SELANGOR	246	17	BEAUTY & SKINCARE
4	NIXODERM BIO-SULFUR WITH SALICYLIC ACID LIQUID...	25.69	KEDAH	41	20	BEAUTY & SKINCARE
5	LAFRE SANITARY PADS PANTYLINER 155MM 20S   LAD...	12.50	NEGERI SEMBILAN	126	21	BEAUTY & SKINCARE
6	COSRX LOW PH GOOD MORNING GEL CLEANSER FOR FAC...	20.44	SELANGOR	57	18	BEAUTY & SKINCARE
7	[READY STOCK] THE GOAT SKINCARE ORGANIC SOAP C...	8.70	PAHANG	89	19	BEAUTY & SKINCARE
8	POND'S SERUM WHIP BRIGHT BEAUTY / PURE BRIGHT ...	14.50	SELANGOR	59	18	BEAUTY & SKINCARE
9	NIVEA LIP PEARL SHINE 4.8G	15.84	SELANGOR	65	15	BEAUTY & SKINCARE


Total rows: 4653  
Total columns: 6

- 🏷️ Category count:
- BEAUTY & SKINCARE: 1279 products
  - STATIONERY: 1083 products
  - HEALTH & WELLNESS: 1031 products
  - HOME APPLIANCES: 537 products
  - HOME & LIVING: 469 products
  - WOMEN'S FASHION: 231 products
  - MOTHER & BABY: 23 products

Group 4 (Most popular):

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	ARON VITAMIN E MOISTURISING CREAM ENRICHED WIT...	7.88	PERLIS	127	27	BEAUTY & SKINCARE
1	MILK WHITENING BODY LOTION SUAVE SKIN CARE MOI...	15.94	SELANGOR	62	25	BEAUTY & SKINCARE
2	SET JRAGAT 3IN1(SABUN. CREAM MALAM DAN SIANG)	4.99	WP KUALA LUMPUR	264	27	BEAUTY & SKINCARE
3	ORIGINAL TOMATO SOAP BY BRILLIANT SKIN	8.99	WP KUALA LUMPUR	214	25	BEAUTY & SKINCARE
4	MENS FACIAL CLEANSER BLACKHEAD REMOVAL FACIAL ...	24.00	SELANGOR	48	24	BEAUTY & SKINCARE
5	DR RASHEL ALOE VERA FACIAL CLEANSER ..	11.30	WP KUALA LUMPUR	57	27	BEAUTY & SKINCARE
6	DASHING 2 IN 1 MUKA DAN BADAN PEMBERSIH 700G F...	21.50	SELANGOR	72	22	BEAUTY & SKINCARE
7	150ML NICOR MEN FACIAL CLEANSER ANTI MITE ACNE...	19.98	JOHOR	232	24	BEAUTY & SKINCARE
8	A BONNE MILK POWER LIGHTENNING LOTION + COLLAG...	22.99	MELAKA	72	22	BEAUTY & SKINCARE
9	HEALTHY SHOP LIPSTICK	27.00	WP KUALA LUMPUR	126	22	BEAUTY & SKINCARE

Total rows: 2783  
Total columns: 6

 Category count:  
- BEAUTY & SKINCARE: 758 products  
- HEALTH & WELLNESS: 636 products  
- STATIONERY: 609 products  
- HOME APPLIANCES: 329 products  
- HOME & LIVING: 287 products  
- WOMEN'S FASHION: 142 products  
- MOTHER & BABY: 22 products

===== Performance =====

Total rows processed: 113,596  
Code Execution time: 13.5979 seconds  
Throughput: 8,353.95 rows per second  
Current memory usage: 0.1762 MB  
Peak memory usage: 0.3121 MB  
CPU usage: 19.0%

Total time for this cell (Including time to display the performance):  
CPU times: user 852 ms, sys: 40.4 ms, total: 893 ms  
Wall time: 14.6 s

Step 6 : Ranking Location Based on Market Performance

In [ ]:

```
%%time

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

df_location_sales = df_price.groupby("Location").agg(
    F.sum("Quantity Sold").alias("Total Quantity Sold"),
    F.avg("Price").alias("Average Price")
)

df_location_sales = df_location_sales.withColumn(
    "Market Performance",
    F.col("Total Quantity Sold") * F.col("Average Price")
)

df_location_sales = df_location_sales.orderBy(F.col("Market Performance").desc())

display(df_location_sales.limit(10).toPandas())
print("\n")

total_rows = df_cleaned.count()

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

execution_time = end_time - start_time
throughput = total_rows / execution_time if execution_time > 0 else 0

print("\n===== 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):")
```

	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

## End of Part 2 Data Optimization