Part1 Data Processing and Cleaning

Prepared by:

Step 1: Install and Impoet 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
import os
import shutil

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 Lazada (Beauty & Skincare).xlsx to Lazada (Beauty & Skincare).xlsx
Saving Lazada (Health & Wellness).xlsx to Lazada (Health & Wellness).xlsx
Saving Lazada (Home & Living).xlsx to Lazada (Home & Living).xlsx
Saving Lazada (Home Appliances).xlsx to Lazada (Home Appliances).xlsx
Saving Lazada (Mother & Baby).xlsx to Lazada (Mother & Baby).xlsx
Saving Lazada (Stationery).xlsx to Lazada (Stationery).xlsx
Saving Lazada (Women's Fashion).xlsx to Lazada (Women's Fashion).xlsx
```

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

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

```
Total DataFrames in fList_pandas: 7
```

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

In [ ]:
```
%%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):")
```

```
Total DataFrames in fList_spark: 7
```

	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: 28325
Total columns: 6
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	HIMALAYA Mentat Tablets 60 (Mind wellness, Enh...	15.00	Wp Kuala Lumpur	45 sold	(8)	Health & Wellness
1	Pil Beauty Original Indonesia | Andalan 2 Strip	5.00	Selangor	56 sold	(11)	Health & Wellness
2	ELECTRAL FORTE GRANULES (20SX6G) - ORS GARAM M...	9.90	Selangor	84 sold	(22)	Health & Wellness
3	[READY STOCK] Spes 1pc Dry Shampoo 5ml 24h Oil...	5.90	Johor	13 sold	(3)	Health & Wellness
4	CALSONATE Calcium Carbonate Capsule 10s (EXP01...	1.54	Melaka	676 sold	(11)	Health & Wellness
5	Super Greens, 1350 mg Per Serving, 120 Capsule...	9.10	Selangor	None	None	Health & Wellness
6	Dyna U Suspension (Peppermint Flavour) 120ml |...	4.00	Selangor	299 sold	(10)	Health & Wellness
7	🔥 Ready Stock 🔥 Travel Pack 25g Maxsure Platinum ...	10.00	Selangor	None	None	Health & Wellness
8	YSP Homecare Macgel Tablet 10 tablets	1.99	Penang	551 sold	(21)	Health & Wellness
9	Ubat Gastrik/ Gastric/ Sakit Perut/ Angin/ Ind...	2.50	Wp Kuala Lumpur	26 sold	(2)	Health & Wellness

```
Total rows: 24382
Total columns: 6
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	80x120/80x160cm Home living room bedroom floor...	7.71	Johor	4.7K sold	(1486)	Home & Living
1	80x120/80x160cm Home living room bedroom floor...	7.79	Johor	860 sold	(235)	Home & Living
2	Cotton Pillow Sleeping Hilton Pillow 1Kg Viral...	6.99	Selangor	121 sold	(23)	Home & Living
3	Sevenland 1pc 70x39cm PE Foam 3D Wall Stickers...	1.80	None	309.9K sold	(4742)	Home & Living
4	MISO Foldable Mosquito Net 1.8 King/1.5 Queen ...	13.34	Selangor	953 sold	(180)	Home & Living
5	Sleeping Hilton Cotton Pillow Hotel Bedding Sl...	12.12	Johor	6 sold	None	Home & Living
6	Leego 50CM X 80CM Carpet Mat Bathroom crystal ...	5.90	Selangor	503 sold	(115)	Home & Living
7	Exclusive Home & Living Sarung Bantal Peluk Be...	4.90	Selangor	106 sold	(24)	Home & Living
8	3pcs/set Astronaut Decoration Cute Model Littl...	3.61	Selangor	206 sold	(113)	Home & Living
9	Cotton Pillow Sleeping Hilton Pillow 1Kg Viral...	8.50	Selangor	148 sold	(34)	Home & Living

```
Total rows: 13376
Total columns: 6
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	ACTIVEONE Home Appliance Household Multifuncti...	31.90	Penang	468 sold	(160)	Home Appliances
1	Teemo Home Appliance Household Multifunctional...	31.90	Penang	112 sold	(39)	Home Appliances
2	DESSINI ITALY 250mL USB Rechargeable Capsule C...	12.50	Selangor	1.6K sold	(611)	Home Appliances
3	GTE Home Appliance Household Multifunctional E...	31.90	Penang	229 sold	(80)	Home Appliances
4	Blender Ready Stock Electric Stainless Steel S...	22.88	Johor	2.0K sold	(528)	Home Appliances
5	2L Electric Meat Grinder Food Chopper Fruits V...	17.99	Pahang	251 sold	(90)	Home Appliances
6	Kettle Stainless Steel 2Liter Electric Kettle ...	14.50	Selangor	9.0K sold	(2286)	Home Appliances
7	Electric Jug Kettle 2L Stainless Steel 2.3L Co...	17.90	Selangor	4.1K sold	(1154)	Home Appliances
8	Portable Turbo Electric Fan 100 Speed Wind adj...	19.60	Selangor	7 sold	(1)	Home Appliances
9	Mafababe Portable Electric Grinder Large Capac...	26.75	Wp Kuala Lumpur	1.0K sold	(308)	Home Appliances

```
Total rows: 13674
Total columns: 6
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	RELAXING MOODS FOR MOTHER & BABY (2XDISC) MU...	54.90	Perak	None	None	Mother & Baby
1	Cotton Breastfeeding Maternity Nursing Cover B...	19.90	Selangor	46 sold	(10)	Mother & Baby
2	Cotton Breastfeeding Nursing Cover Apron Shawl...	8.90	Selangor	802 sold	(233)	Mother & Baby
3	[Malaysia] Breastfeeding Nursing Cover Cotton ...	7.50	Selangor	2.9K sold	(794)	Mother & Baby
4	SummerGlitz 100% Cotton & Cotton Nursing Cover...	24.90	Selangor	1.1K sold	(272)	Mother & Baby
5	KOGGY 5 Pairs Maternity Socks Stokin Pantang F...	11.87	Johor	163 sold	(49)	Mother & Baby
6	Moo Baby Maternity Socks Stokin Pantang Fluffy...	1.90	Perak	5.2K sold	(816)	Mother & Baby
7	[PRE-ORDER] Mothers Baby Coolberry Standard Wr...	49.33	Selangor	None	None	Mother & Baby
8	Breastfeeding Mum Baby Infant Nursing Cover wi...	9.99	Wp Kuala Lumpur	None	None	Mother & Baby
9	Einmilk Baby Cotton Nursing Cover Breastfeedin...	26.90	Johor	25 sold	(9)	Mother & Baby

```
Total rows: 2440
Total columns: 6
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	Deli Direct Liquid Gel Pen Quick Drying 0.5mm ...	0.70	Negeri Sembilan	764 sold	(182)	Stationery
1	Colourful Flexible Pencil Soft Bendable And Tw...	0.39	Perak	508 sold	(5)	Stationery
2	Stationery📐Stationery Set Cartoon Stationery G...	0.98	Perak	275 sold	(4)	Stationery
3	⭐学生卡通中性笔可爱风0.5黑色签字笔⭐New Style School Stationer...	0.49	Perak	2.1K sold	(148)	Stationery
4	*Original* M&G R3/R5 0.5 0.7 Gel Pen- 1 pcs / ...	0.75	Selangor	64.8K sold	(4015)	Stationery
5	12PCS/Set Basics Premium Multicolor Colored Pe...	0.50	Selangor	55.4K sold	(12408)	Stationery
6	（No need to sharpen pencils）Free Rubber 不削铅笔Pe...	0.54	Selangor	8.5K sold	(702)	Stationery
7	⭐学生卡通本创意文具高颜值笔记本子学习用品记事本⭐Notebook Cartoon Diar...	0.29	Perak	3.5K sold	(40)	Stationery
8	Classic Gel Pen 1008 Black Blue Red Ink 0.5mm ...	0.39	Perak	16.2K sold	(492)	Stationery
9	Sanrio Eraser Stationery Non-Dandruff Eraser C...	0.96	Selangor	1.1K sold	(183)	Stationery

```
Total rows: 23195
Total columns: 6
```

	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
```

Step 5 : Combine All DataFrames into One (Data Integration), Checking on Total Numbers of Rows and Columns

In [ ]:

```
%%time

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

df_combined = reduce(lambda df1, df2: df1.unionByName(df2), fList_spark)

display(df_combined.limit(10).toPandas())

total_rows = df_combined.count()
total_columns = len(df_combined.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("\n==================== Performance ====================\n")
print(f"Total rows processed: {total_rows:,}")  # Format with comma separator
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):")
```

	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: 4.8765 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
```

Step 6 : Standardizing Field with Object/String Data Type into Uppercase

In [ ]:

```
%%time

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

for col_name, dtype in df_combined.dtypes:
    if dtype == 'string':
        df_combined = df_combined.withColumn(col_name, F.upper(F.col(col_name)))

display(df_combined.limit(10).toPandas())

total_rows = df_combined.count()
total_columns = len(df_combined.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("\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):")
```

	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: 4.2237 seconds
Throughput: 27,248.92 rows per second
Current memory usage: 0.0679 MB
Peak memory usage: 0.1824 MB
CPU usage: 61.0%

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
```

Step 7 : Converting 'Total Reviews' into Integer Data Type

In [ ]:

```
%%time

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

df_combined = df_combined.withColumn(
    "Total Reviews",
     F.regexp_extract(F.col("Total Reviews"), r"(\d+)", 1).cast("long")
)

display(df_combined.limit(10).toPandas())

total_rows = df_combined.count()
total_columns = len(df_combined.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("\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):")
```

	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
```

Step 8 : Converting 'Quatity Sold' into Integer Data Type

In [ ]:

```
%%time

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

df_combined = df_combined.withColumn(
    "Quantity Sold",
    F.col("Quantity Sold").cast("string")
)

df_combined = df_combined.withColumn(
    "Quantity Sold",
    F.trim(
        F.regexp_replace(F.col("Quantity Sold"), "(?i)sold", ""))
)

df_combined = df_combined.withColumn(
    "Quantity Sold",
    F.when(
        F.col("Quantity Sold").contains("K"),
        (F.regexp_replace(F.col("Quantity Sold"), "K", "").cast("float") * 1000)
    ).otherwise(
        F.regexp_replace(F.col("Quantity Sold"), "[^0-9\\.]", "").cast("float")
    )
)

df_combined = df_combined.withColumn(
    "Quantity Sold",
    F.round(F.col("Quantity Sold")).cast("int")
)

display(df_combined.limit(10).toPandas())

total_rows = df_combined.count()
total_columns = len(df_combined.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("\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):")
```

	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: 6.0%

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
```

Step 9 : Checking and Handling Missing Values by Replacing 0 for Numeric Fields and N/A for String/Object Fields

In [ ]:
```
%%time

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

print("Initial Dataset:")
display(df_combined.limit(10).toPandas())

print("Before Handling Missing Values")
missing_values = df_combined.select([
    F.sum(F.when(F.col(col).isNull(), 1).otherwise(0)).alias(f"{col}_missing")
    for col in df_combined.columns
])
print("Missing Values:")
display(missing_values.toPandas())

df_filled = df_combined

for field in df_filled.schema.fields:
    if isinstance(field.dataType, StringType):
        df_filled = df_filled.withColumn(
            field.name,
            F.when(F.col(field.name).isNull(), F.lit("N/A")).otherwise(F.col(field.name))
        )
    elif isinstance(field.dataType, NumericType):
        df_filled = df_filled.withColumn(
            field.name,
            F.when(F.col(field.name).isNull(), F.lit(0)).otherwise(F.col(field.name))
        )
    else:
        pass

print("\nAfter Handling Missing Values")
missing_values_after = df_filled.select([
    F.sum(F.when(F.col(col).isNull(), 1).otherwise(0)).alias(f"{col}_missing")
    for col in df_filled.columns
])
print("Missing Values:")
display(missing_values_after.toPandas())
```

```
print("\nFinalised Dataset:")
display(df_filled.limit(10).toPandas())

total_rows = df_filled.count()
total_columns = len(df_filled.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("\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):")
```

```
Initial Dataset:
```

	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

```
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
8	❤️88HOME❤️ 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: 115,090
Code Execution time: 18.2981 seconds
Throughput: 6,289.71 rows per second
Current memory usage: 0.1035 MB
Peak memory usage: 0.2372 MB
CPU usage: 12.5%

Total time for this cell (Including time to display the performance):
CPU times: user 785 ms, sys: 22.2 ms, total: 807 ms
Wall time: 19.3 s
```

Step 10 : Checking and Handling Duplicate Rows and Displaying in a View that Arranges Duplicate Rows Together

In [ ]:

```
%%time

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

print("Before Handling Duplicates")

dup_counts = df_filled.groupBy(df_filled.columns).count()
duplicates_only = dup_counts.filter("count > 1").drop("count")
duplicate_rows = df_filled.join(duplicates_only, on=df_filled.columns, how="inner")
duplicate_rows = duplicate_rows.orderBy(df_filled.columns)

num_duplicates = duplicate_rows.count()
print(f"Number of duplicate rows: {num_duplicates}")
display(duplicate_rows.toPandas())
print(f"Total rows: {num_duplicates}")
print(f"Total columns: {len(duplicate_rows.columns)}")

df_cleaned = df_filled.dropDuplicates()

print("\nAfter Handling Duplicates")

duplicate_rows_after = (
    df_cleaned
    .groupBy(df_cleaned.columns)
    .count()
    .filter(F.col("count") > 1)
    .drop("count")
    .orderBy(df_cleaned.columns)
)

num_duplicates_after = duplicate_rows_after.count()
print(f"Number of duplicate rows: {num_duplicates_after}")
display(duplicate_rows_after.toPandas())
print(f"Total rows: {num_duplicates_after}")
print(f"Total columns: {len(duplicate_rows_after.columns)}")

print("\nFinalised Dataset:")
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}\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 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):")
```

```
Before Handling Duplicates
Number of duplicate rows: 2854
```

	Product Name	Price	Location	Quantity Sold	Total Reviews	Category
0	#ARYAN&RAIHAN OOTHING ALOE VERA LIP TREATMENT:...	12.00	WP KUALA LUMPUR	12	4	BEAUTY & SKINCARE
1	#ARYAN&RAIHAN OOTHING ALOE VERA LIP TREATMENT:...	12.00	WP KUALA LUMPUR	12	4	BEAUTY & SKINCARE
2	(ROHTO) HADA-LABO GOKUJUN PREMIUM HYALURONIC...	50.19	JAPAN	0	1	BEAUTY & SKINCARE
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🔥超便宜超便宜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 🥥 ลิปบาล์มน้ำมั...	12.00	SELANGOR	0	0	BEAUTY & SKINCARE
2853	🥥 COCONUT OIL NATURAL LIP BALM 🥥 ลิปบาล์มน้ำมั...	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
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: 113,596
Code Execution time: 65.1819 seconds
Throughput: 1,742.75 rows per second
Current memory usage: 0.1962 MB
Peak memory usage: 2.3469 MB
CPU usage: 13.6%

Total time for this cell (Including time to display the performance):
CPU times: user 1.26 s, sys: 53.7 ms, total: 1.32 s
Wall time: 1min 6s
```

Step 11 : Exporting a Cleaned Excel Data File for Data Optimization

In [ ]:

```
%%time

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

df_cleaned.coalesce(1).write.csv(
    "/content/temp_csv_output",
    header=True,
    mode="overwrite",
    quote='"',
```

```
    escape='"'
)

output_dir = "/content/temp_csv_output"
output_file = [f for f in os.listdir(output_dir) if f.startswith("part-") and f.endswith(".csv")]

if output_file:
    src_path = os.path.join(output_dir, output_file[0])
    final_path = "/content/pyspark_cleaned_dataset.csv"
    shutil.move(src_path, final_path)

    from google.colab import files
    files.download(final_path)
else:
    print("No CSV part file found.")

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):")
```

```
==================== 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
```

End of Part 1 Data Processing and Cleaning