In [1]:
```
## Author: Ashantha Rosary , Goh Boon Xiang , Vithiya
```

In [1]:
```
from pyspark.sql import SparkSession

spark = SparkSession\
        .builder\
        .appName("DataFrameDemo")\
        .getOrCreate()
```

```
24/09/07 19:32:59 WARN Utils: Your hostname, goh1267. resolves to a loopback address: 127.0.1.1; using 10.255.255.254 instead
(on interface lo)
24/09/07 19:32:59 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/09/07 19:33:00 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
```

In [2]:
```
sc = spark.sparkContext
sc.addFile("../de_classes/data_storage/hadoop_file_handler.py")

# Import the HadoopFileHandler class
from hadoop_file_handler import HadoopFileHandler

# Create an instance of HadoopFileHandler
handler = HadoopFileHandler()

# Read raw data from HDFS
df = handler.read_csv('G3_B/data/merged/merged_reviews.csv')
```

```
24/09/07 19:33:03 WARN SparkSession: Using an existing Spark session; only runtime SQL configurations will take effect.
```

In [3]:
```
df.show()
```

```
+----------+------------------------------------+--------------------+------------+---------+
|      Name|                              Review|             SkuInfo|        Date|StarCount|
+----------+------------------------------------+--------------------+------------+---------+
| Angela W.|                Received item in ...|Color Family:Whit...| 08 Oct 2023|        5|
|   HONG Y.|                Why my packing go...|Color Family:Whit...| 13 May 2024|        5|
|Iswaran M.|                The KF94 mask pro...|Color Family:Grey...| 29 Feb 2024|        5|
| Angela W.|                Received items in...|Color Family:Blac...| 02 Jan 2024|        5|
|     A***.|                Finally received ...|Color Family:Whit...| 17 Jan 2024|        4|
|     L***M|                Finally received ...|Color Family:Blac...|4  weeks ago|        5|
|   Jestine|                Order received in...|Color Family:Blac...| 06 Oct 2023|        5|
|     T***m|                Order received in...|Color Family:Whit...| 28 Jun 2024|        5|
| Nazari M.|                Goods received in...|Color Family:Whit...| 30 Jun 2024|        5|
|  ching W.|口罩品质应该没有问题，已经是N次购...|Color Family:Blac...| 24 Nov 2023|        4|
|     L***.|                items safely rece...|Color Family:Blac...| 7  days ago|        5|
|     1***2|                The KF94 mask pro...|Color Family:Whit...| 07 Jun 2024|        5|
| Angela W.|                Received item in ...|Color Family:Blac...| 08 Oct 2023|        5|
|   leow B.|                Good receive in g...|Color Family:Whit...| 02 Oct 2023|        5|
|     L***.|                The goods are rec...|Color Family:Blac...| 20 Apr 2024|        5|
|    lam K.|                Good 👍 👍 👍 👍 ...|Color Family:Blac...| 16 Aug 2023|        5|
|  Chong M.|                repeat purchase a...|Color Family:Whit...|4  weeks ago|        5|
|   Siew M.|                good, fast delivery.|Color Family:Grey...| 10 Jun 2024|        5|
|     s***n|                Item received in ...|Color Family:Whit...| 30 Aug 2023|        5|
|  ching W.|                Item received in ...|Color Family:Grey...| 02 Aug 2023|        5|
+----------+------------------------------------+--------------------+------------+---------+
only showing top 20 rows
```

The 4th row's date is not in proper date format, emojis are present, and the SkuInfo column requires preprocessing.

In [4]:
```
total_rows = df.count()
print(f"Total number of rows: {total_rows}")
```

```
Total number of rows: 5767
```

In [5]:
```
df.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- Review: string (nullable = true)
 |-- SkuInfo: string (nullable = true)
 |-- Date: string (nullable = true)
 |-- StarCount: integer (nullable = true)
```

Data Preprocessing

In [6]:

```
sc.addFile("../de_classes/data_preparation/data_preprocessor.py")
```

In [7]:

```
from data_preprocessor import DataPreprocessor

reference_date = '2024-08-03'  # Reference date in yyyy-MM-dd format(the date we collect data)

preprocessor = DataPreprocessor(df,spark)

 # Remove missing values based on 'Review' column and duplicates based on the entire row
df_cleaned = (preprocessor
                .remove_missing_values(columns=['Review'])
                .remove_duplicates()
                .convert_relative_dates("Date", reference_date)  # Convert 'Date' column to standard date format
                .get_cleaned_data())

# Show the cleaned DataFrame
df_cleaned.printSchema()
df_cleaned.show(truncate=False)
```

```
root
 |-- Name: string (nullable = true)
 |-- Review: string (nullable = true)
 |-- SkuInfo: string (nullable = true)
 |-- Date: date (nullable = true)
 |-- StarCount: integer (nullable = true)

+----------+-------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
---------------+-----------------------------+----------+---------+
|Name      |Review
|SkuInfo                      |Date      |StarCount|
+----------+-------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
---------------+-----------------------------+----------+---------+
|Yan E.    |The goods have been received, thank you seller 😊😊
|Color Family:Black-50PCS     |2023-08-29|5        |
|Thevagi S.|all good
|Color Family:Grey            |2022-05-30|5        |
|s***.     |The quality is not very good, too thin, not genuine.  Considering the price, this is the quality!\n质量不是很好，太
单薄了，不是正版的。 考虑到这个价格，就是这般质量吧！
|color_family:Black           |2022-10-30|3        |
|BC K.     |products packaging not see through, can't see the mask colour inside.
|Color Family:Black           |2022-04-03|4        |
|Jennie 8. |Good quality for the product,,,excellent services\nFrom the seller
|Color Family:Black           |2022-08-01|5        |
|S***.     |轻触并按住剪贴内容即可将其固定。未固定的剪贴内容将于 1 小时之后被删除。欢迎使用 Gboard 剪贴板，您复制的所有文本都会保存到这
里。
|color_family:Black           |2024-01-29|5        |
|Eddie T.  |Good seller fast deliveryGood seller fast deliveryGood seller fast delivery
|color_family:Careion 3D Black|2023-09-18|5        |
|Thi N.    |I think it is good for me. i will use it and review it later for you guys
|color_family:White           |2023-08-31|5        |
|Narainis  |Suka suka suka. Maaf gambar tidak berkaitan tapi produk semua terbaik
|color_family:Hitam           |2023-04-05|5        |
|Khor S.   |thanks received
|color_family:Black           |2022-12-21|5        |
|Tham      |very thin
|Color Family:Black-50PCS     |2022-11-27|5        |
|*******898|It's unbelievable that seller sent  me short of 1 item out of only 20 I ordered.. I can only assume that it's done
intentionally. The mask is only 3ply and not 4ply as advertised and felt cheated by seller. I forgo my claim as the process is
troublesome and its cost is minimal. The courier service really suckered and Lazada should ensure that buyers deserve better de
livery service.|Color Family:Purple          |2022-05-04|3        |
|Yap B.    |Fast shipping, will repurchase again.  Thank you
|color_family:Black           |2023-06-05|5        |
|W***.     |Selamat sampai thx seller\ncuma nipis sikit\nsesuai dg harga
|Color Family:Hitam           |2022-05-28|5        |
|N***.     |mask cukup seperti yang di pesan...harga murah boleh beli lagi
|Color:Grey                   |2022-04-01|5        |
|K***.     |pelik......kite order mask lain, tgok2 mask lain yg sampai....SGT2 kecewa la...walaupun murah...mask yg jenis lain
yg dibagi tu, 50pcs 1.50 tp nipis giler...sy nk mask lain awak eh
|Color:Headloop Purple        |2022-04-10|5        |
|Theresa T.|The item just received today which packed neat.  The mask same as per advertised and comfortable to wear. Thank you
Lazada for prompt delivery.
|Color Family:White-50PCS     |2023-11-26|5        |
|1***1     |I ordered on sept(50pcs) to check the size and design.. all were good. so I re ordered 100pcs. end up different des
ign at nose bridge.dissappointed.
|Color Family:Black-50PCS     |2022-11-14|5        |
|Lun K.    |First time buy it, feel good
|Color:White                  |2022-04-25|5        |
|L***.     |This mask is big enough for man's face.\nIt is comfortable, good price too.\nThank you seller and courier guy.
|Color Family:Grey-50PCS      |2022-09-12|5        |
+----------+------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
---------------+-----------------------------+----------+---------+
only showing top 20 rows
```

Proof of successful changing "4 weeks ago" to exact date

In [8]:

```python
# Register the DataFrame as a SQL temporary view
df_cleaned.createOrReplaceTempView("reviews")

result_df = spark.sql("SELECT * FROM reviews WHERE Name = 'Chong M.'")

result_df.show(truncate=True)
```

```
+--------+--------------------+--------------------+----------+---------+
|    Name|              Review|             SkuInfo|      Date|StarCount|
+--------+--------------------+--------------------+----------+---------+
|Chong M.|repeat purchase a...|Color Family:Whit...|2024-07-06|        5|
+--------+--------------------+--------------------+----------+---------+
```

In [9]:

```
total_rows = df_cleaned.count()
print(f"Total number of rows after data preprocessing: {total_rows}")
```

```
Total number of rows after data preprocessing: 3450
```

In [10]:

```
from pyspark.sql.functions import asc

# Group by StarCount and count the occurrences, then order by StarCount in ascending order
df_cleaned.groupBy("StarCount").count().orderBy(asc("StarCount")).show()
```

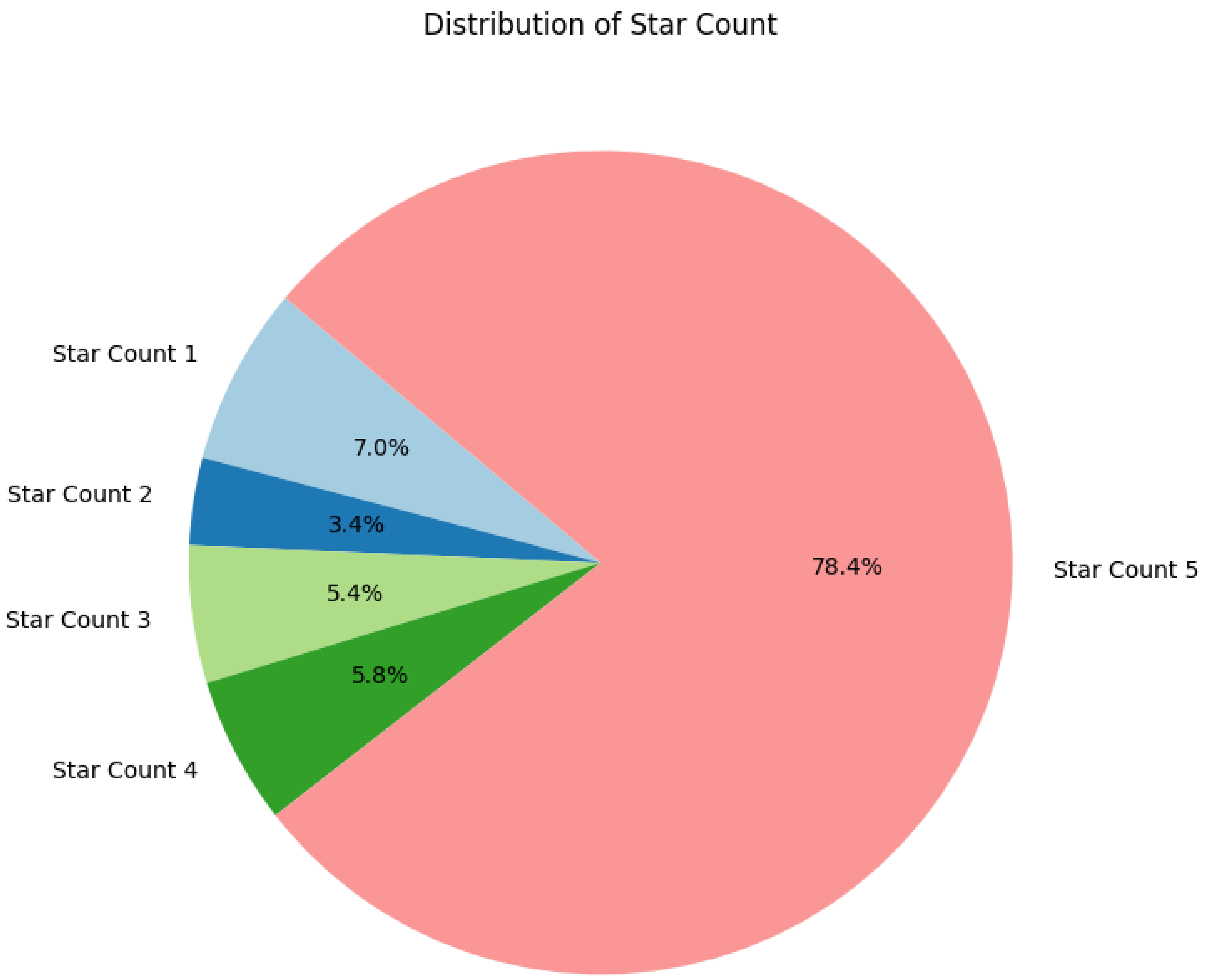
```
+---------+-----+
|StarCount|count|
+---------+-----+
|        1|  242|
|        2|  118|
|        3|  186|
|        4|  200|
|        5| 2704|
+---------+-----+
```

In [11]:

```
sc.addFile("../de_classes/data_visualisation/starCount.py")
```

In [12]:

```
# Import the custom visualizer class
from starCount import StarCountVisualizer

# Initialize the StarCountVisualizer with the cleaned Spark DataFrame
visualizer = StarCountVisualizer(df_cleaned)

visualizer.prepare_data()
visualizer.plot_pie_chart()
```


Distribution of Star Count
Star Count 1
7.0%
Star Count 2
3.4%
Star Count 3
5.4%
Star Count 4
5.8%
78.4%
Star Count 5

In [13]:

```
distinct_skuinfo = spark.sql("SELECT DISTINCT SkuInfo FROM reviews")
distinct_skuinfo.show(truncate=False)
```

```
+---------------------------------+
|SkuInfo                          |
+---------------------------------+
|color_family:Orange              |
|Color Family:Hitam               |
|Color Family:Green               |
|Color:Black                      |
|color_family:Headloop Navy Blue  |
|color_family:Headloop Light Blue |
|Color Family:TiffanyPattern20PCS |
|Color Family:Light Blue          |
|color_family:Headloop Light Green|
|color_family:RedPattern20PCS     |
|Color:Headloop Pink              |
|color_family:Headloop Ombre Gradi|
|Color Family:Rose Gold-50PCS     |
|Color Family:White Green         |
|Color Family:Black               |
|Color:Rose Gold                  |
|color_family:BlackPattern20PCS   |
|Color:Light blue                 |
|Color Family:Kelabu              |
|color_family:Random Pattern      |
+---------------------------------+
only showing top 20 rows
```

In [14]:

```
distinct_count = spark.sql("SELECT COUNT(DISTINCT SkuInfo) AS distinct_count FROM reviews")
distinct_count.show()
```

```
+--------------+
|distinct_count|
+--------------+
|           114|
+--------------+
```

Storing in Redis & Retrieve from Redis

In [15]:

```
sc.addFile("../de_classes/data_storage/redis_handler.py")
```

In [16]:

```
from redis_handler import RedisHandler

redis_handler = RedisHandler(host='localhost', port=6379, db=0)

# Store the DataFrame in Redis
num_rows = redis_handler.store_dataframe(df_cleaned)

# Load the data back from Redis
loaded_data = redis_handler.load_data(num_rows)

# Convert the loaded data to a DataFrame
df_loaded = redis_handler.convert_to_dataframe(loaded_data, spark)
df_loaded.show(truncate=False)
```

```
Data stored in Redis successfully.
+----------+--------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
---------------+-----------------------------+----------+---------+
|Name      |Review
|SkuInfo                      |Date      |StarCount|
+----------+--------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
---------------+-----------------------------+----------+---------+
|Yan E.    |The goods have been received, thank you seller 😊😊
|Color Family:Black-50PCS     |2023-08-29|5        |
|Thevagi S.|all good
|Color Family:Grey            |2022-05-30|5        |
|s***.     |The quality is not very good, too thin, not genuine.  Considering the price, this is the quality!\n质量不是很好，太
单薄了，不是正版的。 考虑到这个价格，就是这般质量吧！
|color_family:Black           |2022-10-30|3        |
|BC K.     |products packaging not see through, can't see the mask colour inside.
|Color Family:Black           |2022-04-03|4        |
|Jennie 8. |Good quality for the product,,,excellent services\nFrom the seller
|Color Family:Black           |2022-08-01|5        |
|S***.     |轻触并按住剪贴内容即可将其固定。未固定的剪贴内容将于 1 小时之后被删除。欢迎使用 Gboard 剪贴板，您复制的所有文本都会保存到这
里。
|color_family:Black           |2024-01-29|5        |
|Eddie T.  |Good seller fast deliveryGood seller fast deliveryGood seller fast delivery
|color_family:Careion 3D Black|2023-09-18|5        |
|Thi N.    |I think it is good for me. i will use it and review it later for you guys
|color_family:White           |2023-08-31|5        |
|Narainis  |Suka suka suka. Maaf gambar tidak berkaitan tapi produk semua terbaik
|color_family:Hitam           |2023-04-05|5        |
|Khor S.   |thanks received
|color_family:Black           |2022-12-21|5        |
|Tham      |very thin
|Color Family:Black-50PCS     |2022-11-27|5        |
|*******898|It's unbelievable that seller sent  me short of 1 item out of only 20 I ordered.. I can only assume that it's done
intentionally. The mask is only 3ply and not 4ply as advertised and felt cheated by seller. I forgo my claim as the process is
troublesome and its cost is minimal. The courier service really suckered and Lazada should ensure that buyers deserve better de
livery service.|Color Family:Purple          |2022-05-04|3        |
|Yap B.    |Fast shipping, will repurchase again.  Thank you
|color_family:Black           |2023-06-05|5        |
|W***.     |Selamat sampai thx seller\ncuma nipis sikit\nsesuai dg harga
|Color Family:Hitam           |2022-05-28|5         |
|N***.     |mask cukup seperti yang di pesan...harga murah boleh beli lagi
|Color:Grey                   |2022-04-01|5         |
|K***.     |pelik......kite order mask lain, tgok2 mask lain yg sampai....SGT2 kecewa la...walaupun murah...mask yg jenis lain
yg dibagi tu, 50pcs 1.50 tp nipis giler...sy nk mask lain awak eh
|Color:Headloop Purple        |2022-04-10|5         |
|Theresa T.|The item just received today which packed neat.  The mask same as per advertised and comfortable to wear. Thank you
Lazada for prompt delivery.
|Color Family:White-50PCS     |2023-11-26|5         |
|1***1     |I ordered on sept(50pcs) to check the size and design.. all were good. so I re ordered 100pcs. end up different des
ign at nose bridge.dissappointed.
|Color Family:Black-50PCS     |2022-11-14|5         |
|Lun K.    |First time buy it, feel good
|Color:White                  |2022-04-25|5         |
|L***.     |This mask is big enough for man's face.\nIt is comfortable, good price too.\nThank you seller and courier guy.
|Color Family:Grey-50PCS      |2022-09-12|5         |
+----------+-------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
---------------+-----------------------------+----------+---------+
only showing top 20 rows
```

SkuInfo Preprocessing

In [17]:
```python
preprocessor = DataPreprocessor(df_loaded, spark)

df_SkuInfo_Processing = (preprocessor.convert_to_lowercase(columns=['SkuInfo'])
                          .remove_punctuation(columns=['SkuInfo'])
                          .remove_color_family_words(columns=['SkuInfo'])
                          .get_cleaned_data())

df_SkuInfo_Processing.show(truncate = False)
```

```
+----------+-------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------------------
---------------+----------------+----------+---------+
|Name      |Review
|SkuInfo         |Date      |StarCount|
+----------+-------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------------------------------------
---------------+----------------+----------+---------+
|Yan E.    |The goods have been received, thank you seller 😊 😊
|black 50pcs     |2023-08-29|5        |
|Thevagi S.|all good
|grey            |2022-05-30|5        |
|s***.     |The quality is not very good, too thin, not genuine.  Considering the price, this is the quality!\n质量不是很好，太
单薄了，不是正版的。 考虑到这个价格，就是这般质量吧！
|black           |2022-10-30|3        |
|BC K.     |products packaging not see through, can't see the mask colour inside.
|black           |2022-04-03|4        |
|Jennie 8. |Good quality for the product,,,excellent services\nFrom the seller
|black           |2022-08-01|5        |
|S***.     |轻触并按住剪贴内容即可将其固定。未固定的剪贴内容将于 1 小时之后被删除。欢迎使用 Gboard 剪贴板，您复制的所有文本都会保存到这
里。
|black           |2024-01-29|5        |
|Eddie T.  |Good seller fast deliveryGood seller fast deliveryGood seller fast delivery
|careion 3d black|2023-09-18|5        |
|Thi N.    |I think it is good for me. i will use it and review it later for you guys
|white           |2023-08-31|5        |
|Narainis  |Suka suka suka. Maaf gambar tidak berkaitan tapi produk semua terbaik
|hitam           |2023-04-05|5        |
|Khor S.   |thanks received
|black           |2022-12-21|5        |
|Tham      |very thin
|black 50pcs     |2022-11-27|5        |
|*******898|It's unbelievable that seller sent  me short of 1 item out of only 20 I ordered.. I can only assume that it's done
intentionally. The mask is only 3ply and not 4ply as advertised and felt cheated by seller. I forgo my claim as the process is
troublesome and its cost is minimal. The courier service really suckered and Lazada should ensure that buyers deserve better de
livery service.|purple          |2022-05-04|3        |
|Yap B.    |Fast shipping, will repurchase again.  Thank you
|black           |2023-06-05|5        |
|W***.     |Selamat sampai thx seller\ncuma nipis sikit\nsesuai dg harga
```

```
|hitam           |2022-05-28|5        |
|N***.     |mask cukup seperti yang di pesan...harga murah boleh beli lagi
|grey            |2022-04-01|5        |
|K***.     |pelik......kite order mask lain, tgok2 mask lain yg sampai....SGT2 kecewa la...walaupun murah...mask yg jenis lain
yg dibagi tu, 50pcs 1.50 tp nipis giler...sy nk mask lain awak eh
|headloop purple |2022-04-10|5        |
|Theresa T.|The item just received today which packed neat.  The mask same as per advertised and comfortable to wear. Thank you
Lazada for prompt delivery.
|white 50pcs     |2023-11-26|5        |
|1***1     |I ordered on sept(50pcs) to check the size and design.. all were good. so I re ordered 100pcs. end up different des
ign at nose bridge.dissappointed.
|black 50pcs     |2022-11-14|5        |
|Lun K.    |First time buy it, feel good
|white           |2022-04-25|5        |
|L***.     |This mask is big enough for man's face.\nIt is comfortable, good price too.\nThank you seller and courier guy.
|grey 50pcs      |2022-09-12|5        |
+----------+-------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------
---------------+----------------+----------+---------+
only showing top 20 rows
```

In [18]:

```python
df_SkuInfo_Processing.createOrReplaceTempView("sku")
distinct_skuinfo = spark.sql("SELECT DISTINCT SkuInfo FROM sku")
distinct_skuinfo.show(n=distinct_skuinfo.count(),truncate=False)
```

```
+--------------------+
|SkuInfo             |
+--------------------+
|grey 50pcs          |
|dark brown          |
|headloop darkblue   |
|grey                |
|green               |
|merah jambu         |
|ombre gradient      |
|headloop black      |
|blackpattern20pcs   |
|headloop ombre gradi|
|black 50pcs         |
|hitam               |
|dark blue 50pcs     |
|headloop white      |
|purple              |
|light brown         |
|white               |
|headloop grey       |
|whitepattern20pcs   |
|pink                |
|red                 |
|careion 3d white    |
|headloop purple     |
|dark blue           |
|headloop light blue |
|light blue          |
|navy blue           |
|putih               |
|dark salmon         |
|black               |
|careion 3d black    |
|matcha 50pcs        |
|white 50pcs         |
|headloop orange     |
|rose gold 50pcs     |
|light green         |
|white green         |
|dark red            |
|kid random pattern  |
|kid ramdom pattern  |
|orange              |
|greypattern20pcs    |
|yellow              |
|magenta             |
|headloop pink       |
|biru muda           |
|headloop light green|
|headloop yellow     |
|purplepattern20pcs  |
|headloop navy blue  |
|random pattern      |
|headloop light brown|
|redpattern20pcs     |
|matcha              |
|headloop brown      |
|kelabu              |
|tiffanypattern20pcs |
|coklat gelap        |
|headloop red        |
|rose gold           |
|kuning              |
+--------------------+
```

Translation

Please skip this step as we have already save it into parquet. Please run the code to load the data from parquet file

In [19]:

```
#please install in VM
#pip install googletrans==4.0.0-rc1
```

In [20]:

```
# from googletrans import Translator
# preprocessor = DataPreprocessor(df_SkuInfo_Processing, spark)
# df_translated_to_English = (preprocessor.translate_column(columns=['Name', 'Review','SkuInfo'])
#                              .get_cleaned_data())
# df_translated_to_English.show(truncate = True)
```

Save as parquet

In [21]:

```
# # Save the DataFrame as a Parquet file
# handler = HadoopFileHandler()
# handler.write_parquet(df_translated_to_English,"data/translated/df_translated_to_English_parquet")
```

Load from parquet into dataframe

In [22]:

```
# Path to your Parquet file
parquet_file_path = "G3_B/data/translated/df_translated_to_English_parquet"

# Load the Parquet file into a DataFrame
df_translated = handler.read_parquet(parquet_file_path)

# Show the DataFrame
df_translated.show(truncate=True)
```

```
+----------+--------------------+----------------+----------+---------+
|      Name|              Review|         SkuInfo|      Date|StarCount|
+----------+--------------------+----------------+----------+---------+
|    Yan E.|The goods have be...|     black 50pcs|2023-08-29|        5|
|Thevagi S.|            all good|            grey|2022-05-30|        5|
|     s***.|The Quality is no...|           black|2022-10-30|        3|
|     BC K.|products packagin...|           black|2022-04-03|        4|
| Jennie 8.|Good quality for ...|           black|2022-08-01|        5|
|     S***.|Touch and hold th...|           black|2024-01-29|        5|
|  Eddie T.|Good seller fast ...|careion 3d black|2023-09-18|        5|
|    Thi N.|I think it is goo...|           white|2023-08-31|        5|
|  Narainis|Likes to love.Sor...|           black|2023-04-05|        5|
|   Khor S.|     thanks received|           black|2022-12-21|        5|
|    Greedy|           very thin|     black 50pcs|2022-11-27|        5|
|*******898|It's unbelievable...|          purple|2022-05-04|        3|
|    Yap B.|Fast shipping, wi...|           black|2023-06-05|        5|
|     W***.|Happy to Thx Sell...|           black|2022-05-28|        5|
|     N***.|Mask is enough as...|            grey|2022-04-01|        5|
|     K***.|Strange ...... se...| headloop purple|2022-04-10|        5|
|Theresa T.|The item just rec...|     white 50pcs|2023-11-26|        5|
|     1***1|I ordered on sept...|     black 50pcs|2022-11-14|        5|
|    Lun K.|First time buy it...|           white|2022-04-25|        5|
|     L***.|This mask is big ...|      grey 50pcs|2022-09-12|        5|
+----------+--------------------+----------------+----------+---------+
only showing top 20 rows
```

Lowercase, No number, Abbrieviation, No punctuations, Emoji Handling

In [23]:

```
sc.addFile("/home/student/G3_B/de_classes/data_preparation/emoji_handler.py")
```

In [24]:

```
from emoji_handler import EmojiHandler

# Load the emoji dictionary
import pickle
emoji_dict = pickle.load(open('/home/student/G3_B/data/merged/Emoji_Dict.p', 'rb'))
emoji_handler = EmojiHandler(emoji_dict)

# Get the UDF for emoji mapping
```

```python
replace_emojis_udf = emoji_handler.get_replace_emojis_udf()
preprocessor = DataPreprocessor(df_translated, spark)

df_Text_Processing = (preprocessor.convert_to_lowercase(columns=['Review', 'SkuInfo'])
                      .remove_words_with_numbers(columns=['Review'])
                      .replace_with_custom_dict(column='Review', dict_path="G3_B/data/merged/dict.csv")
                      .remove_punctuation(columns=['Review'])
                      .get_cleaned_data())

# Apply the emoji mapping to the 'Review' column
df_Text_Processing = df_Text_Processing.withColumn('Review', replace_emojis_udf(df_Text_Processing['Review']))
df_Text_Processing.show(truncate=True)
```

```
+----------+--------------------+----------------+----------+---------+
|      Name|              Review|         SkuInfo|      Date|StarCount|
+----------+--------------------+----------------+----------+---------+
|    Yan E.|the goods have be...|     black 50pcs|2023-08-29|        5|
|Thevagi S.|            all good|            grey|2022-05-30|        5|
|     s***.|the quality is no...|           black|2022-10-30|        3|
|     BC K.|products packagin...|           black|2022-04-03|        4|
| Jennie 8.|good quality for ...|           black|2022-08-01|        5|
|     S***.|touch and hold th...|           black|2024-01-29|        5|
|  Eddie T.|good seller fast ...|careion 3d black|2023-09-18|        5|
|    Thi N.|i think it is goo...|           white|2023-08-31|        5|
|  Narainis|likes to love sor...|           black|2023-04-05|        5|
|   Khor S.|     thanks received|           black|2022-12-21|        5|
|    Greedy|           very thin|     black 50pcs|2022-11-27|        5|
|*******898|it s unbelievable...|          purple|2022-05-04|        3|
|    Yap B.|fast shipping  wi...|           black|2023-06-05|        5|
|     W***.|happy to thanks s...|           black|2022-05-28|        5|
|     N***.|mask is enough as...|            grey|2022-04-01|        5|
|     K***.|strange        se...| headloop purple|2022-04-10|        5|
|Theresa T.|the item just rec...|     white 50pcs|2023-11-26|        5|
|     1***1|i ordered on sept...|     black 50pcs|2022-11-14|        5|
|    Lun K.|first time buy it...|           white|2022-04-25|        5|
|     L***.|this mask is big ...|      grey 50pcs|2022-09-12|        5|
+----------+--------------------+----------------+----------+---------+
only showing top 20 rows
```

In [25]:

```
df_Text_Processing.count()
```

Out[25]:

```
3450
```

Trim, Remove empty rows, Remove stop words

In [26]:

```
preprocessor = DataPreprocessor(df_Text_Processing, spark)

df_Text_Processing2 = (preprocessor.trim_whitespace(columns=['Name','Review', 'SkuInfo'])
                      .remove_empty_and_whitespace_rows(columns =['Review'])
                      .remove_stop_words(column ='Review')
                      .get_cleaned_data())
print(f"Number of rows: {df_Text_Processing2.count()}")
df_Text_Processing2.show(truncate = True)
```

```
Number of rows: 3437
+----------+--------------------+----------------+----------+---------+
|      Name|              Review|         SkuInfo|      Date|StarCount|
+----------+--------------------+----------------+----------+---------+
|    Yan E.|goods have been r...|     black 50pcs|2023-08-29|        5|
|Thevagi S.|            all good|            grey|2022-05-30|        5|
|     s***.|quality not very ...|           black|2022-10-30|        3|
|     BC K.|products packagin...|           black|2022-04-03|        4|
| Jennie 8.|good quality prod...|           black|2022-08-01|        5|
|     S***.|touch hold clippe...|           black|2024-01-29|        5|
|  Eddie T.|good seller fast ...|careion 3d black|2023-09-18|        5|
|    Thi N.|i think good me i...|           white|2023-08-31|        5|
|  Narainis|likes love sorry ...|           black|2023-04-05|        5|
|   Khor S.|     thanks received|           black|2022-12-21|        5|
|    Greedy|           very thin|     black 50pcs|2022-11-27|        5|
|*******898|s unbelievable se...|          purple|2022-05-04|        3|
|    Yap B.|fast shipping wil...|           black|2023-06-05|        5|
|     W***.|happy thanks sell...|           black|2022-05-28|        5|
|     N***.|mask enough messa...|            grey|2022-04-01|        5|
|     K***.|strange see anoth...| headloop purple|2022-04-10|        5|
|Theresa T.|item just receive...|     white 50pcs|2023-11-26|        5|
|     1***1|i ordered sept ch...|     black 50pcs|2022-11-14|        5|
|    Lun K.|first time buy fe...|           white|2022-04-25|        5|
|     L***.|mask big enough m...|      grey 50pcs|2022-09-12|        5|
+----------+--------------------+----------------+----------+---------+
only showing top 20 rows
```

In [27]:
```
df_Text_Processing2.createOrReplaceTempView("ProcessedReviews")
```

In [28]:
```
distinct_count = spark.sql("SELECT COUNT(DISTINCT SkuInfo) AS distinct_count FROM ProcessedReviews")
distinct_count.show()
```

```
[Stage 73:================================================>         (5 + 1) / 6]
+--------------+
|distinct_count|
+--------------+
|            55|
+--------------+
```

In [29]:

```
distinct_skuinfo = spark.sql("SELECT DISTINCT SkuInfo FROM ProcessedReviews")
distinct_skuinfo.show(n=distinct_skuinfo.count(), truncate=False)
```

```
+------------------------------------------------------+
|SkuInfo                                               |
+------------------------------------------------------+
|grey 50pcs                                            |
|dark brown                                            |
|headloop darkblue                                     |
|grey                                                  |
|green                                                 |
|'translator' object has no attribute 'raise_exception'|
|headloop black                                        |
|headloop ombre gradi                                  |
|black 50pcs                                           |
|dark blue 50pcs                                       |
|headloop white                                        |
|purple                                                |
|light brown                                           |
|white                                                 |
|headloop grey                                         |
|whitepattern20pcs                                     |
|pink                                                  |
|red                                                   |
|careion 3d white                                      |
|headloop purple                                       |
|headloop light blue                                   |
|light blue                                            |
|headloop light green                                  |
|navy blue                                             |
|headloop yellow                                       |
|dark salmon                                           |
|black                                                 |
|careion 3d black                                      |
|gradient shadow                                       |
|matcha 50pcs                                          |
|white 50pcs                                           |
|rose gold 50pcs                                       |
|light green                                           |
|white green                                           |
|dark red                                              |
|kid random pattern                                    |
|kid ramdom pattern                                    |
|orange                                                |
|greypattern20pcs                                      |
|magenta                                               |
|purplepattern20pcs                                    |
|headloop navy blue                                    |
|yellow                                                |
|rose gold                                             |
|headloop pink                                         |
|dark blue                                             |
|random pattern                                        |
|headloop orange                                       |
|tiffanypattern20pcs                                   |
|blackpattern20pcs                                     |
|gray                                                  |
|headloop brown                                        |
|headloop light brown                                  |
|redpattern20pcs                                       |
|matcha                                                |
+------------------------------------------------------+
```

Drop row with 'translator' object has no attribute 'raise_exception' AS the SkuInfo

In [30]:

```
# Filter out rows where SkuInfo is the problematic string
problematic_value = "'translator' object has no attribute 'raise_exception'"
filtered_df = df_Text_Processing2.filter(df_Text_Processing2.SkuInfo != problematic_value)

filtered_df.createOrReplaceTempView("ProcessedReviews")

print(f"Number of rows after filtering: {filtered_df.count()}")
filtered_df.show(truncate=True)
```

```
Number of rows after filtering: 2139
+----------+--------------------+----------------+----------+---------+
|      Name|              Review|         SkuInfo|      Date|StarCount|
+----------+--------------------+----------------+----------+---------+
|    Yan E.|goods have been r...|     black 50pcs|2023-08-29|        5|
|Thevagi S.|            all good|            grey|2022-05-30|        5|
|     s***.|quality not very ...|           black|2022-10-30|        3|
|     BC K.|products packagin...|           black|2022-04-03|        4|
| Jennie 8.|good quality prod...|           black|2022-08-01|        5|
|     S***.|touch hold clippe...|           black|2024-01-29|        5|
|  Eddie T.|good seller fast ...|careion 3d black|2023-09-18|        5|
|    Thi N.|i think good me i...|           white|2023-08-31|        5|
|  Narainis|likes love sorry ...|           black|2023-04-05|        5|
|   Khor S.|     thanks received|           black|2022-12-21|        5|
|    Greedy|           very thin|     black 50pcs|2022-11-27|        5|
|*******898|s unbelievable se...|          purple|2022-05-04|        3|
|    Yap B.|fast shipping wil...|           black|2023-06-05|        5|
|     W***.|happy thanks sell...|           black|2022-05-28|        5|
|     N***.|mask enough messa...|            grey|2022-04-01|        5|
|     K***.|strange see anoth...| headloop purple|2022-04-10|        5|
|Theresa T.|item just receive...|     white 50pcs|2023-11-26|        5|
|     1***1|i ordered sept ch...|     black 50pcs|2022-11-14|        5|
|    Lun K.|first time buy fe...|           white|2022-04-25|        5|
|     L***.|mask big enough m...|      grey 50pcs|2022-09-12|        5|
+----------+--------------------+----------------+----------+---------+
only showing top 20 rows
```

In [31]:

```
# Get all distinct SkuInfo values after filtering
distinct_skuinfo = filtered_df.select("SkuInfo").distinct()
distinct_skuinfo.show(n=distinct_skuinfo.count(), truncate=False)
```

```
+--------------------+
|SkuInfo             |
+--------------------+
|grey 50pcs          |
|dark brown          |
|headloop darkblue   |
|grey                |
|green               |
|headloop black      |
|headloop ombre gradi|
|black 50pcs         |
|dark blue 50pcs     |
|headloop white      |
|purple              |
|light brown         |
|white               |
|headloop grey       |
|whitepattern20pcs   |
|pink                |
|red                 |
|careion 3d white    |
|headloop purple     |
|headloop light blue |
|light blue          |
|headloop light green|
|navy blue           |
|headloop yellow     |
|dark salmon         |
|black               |
|careion 3d black    |
|gradient shadow     |
|matcha 50pcs        |
|white 50pcs         |
|rose gold 50pcs     |
|light green         |
|white green         |
|dark red            |
|kid random pattern  |
|kid ramdom pattern  |
|orange              |
|greypattern20pcs    |
|magenta             |
|purplepattern20pcs  |
|headloop navy blue  |
|yellow              |
|rose gold           |
|headloop pink       |
|dark blue           |
|random pattern      |
|headloop orange     |
|tiffanypattern20pcs |
|blackpattern20pcs   |
|gray                |
|headloop brown      |
|headloop light brown|
|redpattern20pcs     |
|matcha              |
+--------------------+
```

Neo4J Storing

In [32]:

```
sc.addFile("/home/student/G3_B/de_classes/data_storage/neo4j_file_handler.py")
```

In [33]:

```
from neo4j_file_handler import Neo4jHandler


# Initialize Neo4jHandler
uri = "neo4j+s://1e220ea1.databases.neo4j.io"
user = "neo4j"
password = "5OGMEMwPRK_NoMZEBHz-pcFP_iyIMwRXxbpplFAD94E"
neo4j_handler = Neo4jHandler(uri, user, password)

# Clear database
neo4j_handler.clear_database()

try:
    data_list = [row.asDict() for row in filtered_df.collect()]
    # Create product nodes and relationships
    neo4j_handler.create_product_nodes_and_relationships(data_list)
except Exception as e:
    print(f"Error processing DataFrame: {e}")
```

```
Successfully connected to Neo4j!
Database cleared successfully.
```

```
Batch 1 processed successfully.
Batch 2 processed successfully.
Batch 3 processed successfully.
Total Product nodes: 54
Total Review nodes: 2139
```

Neo4J Retrieving & Load

In [34]:

```
# Load reviews to Spark DataFrame
try:
    df = neo4j_handler.load_reviews_to_dataframe(spark)
    df.show(100)
except Exception as e:
    print(f"Error loading reviews: {e}")

# Close the connection
neo4j_handler.close()
```

```
+--------------------+--------------------+-----------+----------+---------+
|                Name|              Review|    SkuInfo|      Date|StarCount|
+--------------------+--------------------+-----------+----------+---------+
|               Nasih|             awesome|black 50pcs|2022-07-27|        5|
|          Abinash M.|                  ok|black 50pcs|2022-09-28|        5|
|              Loh W.|   great design love|black 50pcs|2024-06-10|        5|
|               Md A.|        good product|black 50pcs|2023-07-15|        5|
|           Garlic M.|doesnt match vide...|black 50pcs|2022-07-25|        1|
|            NorHa S.|black mask there ...|black 50pcs|2022-08-16|        5|
|               n***i|received good con...|black 50pcs|2022-06-28|        5|
|          THERESA H.|fast delivery but...|black 50pcs|2022-10-03|        5|
|          *******896|short three packs...|black 50pcs|2023-05-12|        4|
|            Chris C.|got packing no so...|black 50pcs|2023-09-26|        5|
|              Yan E.|goods have been r...|black 50pcs|2023-08-29|        5|
|              Greedy|           very thin|black 50pcs|2022-11-27|        5|
|               1***1|i ordered sept ch...|black 50pcs|2022-11-14|        5|
|               T***.|good service fast...|black 50pcs|2024-03-03|        5|
|              AgnesQ|inner layer mask ...|black 50pcs|2023-02-27|        5|
|              tan F.|no good quality l...|black 50pcs|2022-07-29|        1|
|             Arse A.|congratulations u...|black 50pcs|2022-07-06|        5|
|               L***.|nice mask fast de...|black 50pcs|2022-07-23|        5|
|               GO A.|not ply made chin...|black 50pcs|2023-02-25|        1|
|             Tang L.|everything good f...|black 50pcs|2023-06-09|        5|
|               I***.|received good con...|black 50pcs|2022-12-12|        5|
|          Theresa T.|thanks seller goo...|black 50pcs|2023-09-06|        5|
|           Joshua L.|good comfy worth ...|black 50pcs|2022-09-08|        5|
|               K***.|fast delivery its...|black 50pcs|2023-06-09|        5|
|            Fanny L.|           very good|black 50pcs|2022-06-22|        5|
|             Rosnani|           tq seller|black 50pcs|2022-08-13|        4|
|             Khoo S.|fast delivery cor...|black 50pcs|2024-01-25|        5|
|             Sharvin|    everything great|black 50pcs|2023-01-17|        5|
|             Iris G.|       good products|black 50pcs|2022-08-22|        5|
|              Low L.|    very low quality|black 50pcs|2022-07-15|        1|
|               D***l|product received ...|black 50pcs|2024-03-18|        2|
|         Muhammad S.|        s been while|black 50pcs|2022-09-22|        5|
|               0***6|received yesterda...|black 50pcs|2022-07-25|        5|
|               J***.|good wear so comf...|black 50pcs|2022-08-20|        5|
|              Flyntz|comfortable price...|black 50pcs|2023-07-14|        5|
|               ****0|mask ply bukan pl...|black 50pcs|2023-04-11|        2|
|          Gregory C.|quality mask wort...|black 50pcs|2022-12-30|        5|
|               V***.|very fast deliver...|black 50pcs|2022-07-09|        5|
|           Adrian S.|black mask there ...|black 50pcs|2022-04-11|         5|
|               1***9|yeppi yeseseseses...|black 50pcs|2024-01-15|         5|
|            Kangsika|              thanks|black 50pcs|2022-08-08|         5|
|               l***.|alhamdulillah ite...|black 50pcs|2022-07-29|         5|
|              Yap C.|mass mask very go...|black 50pcs|2023-05-06|         5|
|               1***9|stick quickly rec...|black 50pcs|2023-01-03|         5|
|               Imahh|           tq saller|black 50pcs|2023-04-14|         5|
|         You know P.|           thank you|black 50pcs|2022-09-26|         5|
|            River S.|              thanks|black 50pcs|2023-09-30|         5|
|        christina L.|delivered very fa...|black 50pcs|2022-12-29|         5|
|'Translator' obje...|same photo but ve...|black 50pcs|2022-07-07|         3|
|               W***.|i received goods ...|black 50pcs|2023-10-18|         5|
|          Theresa T.|received good con...|black 50pcs|2023-05-11|         5|
|               ng J.|product delivery ...|black 50pcs|2022-10-18|         4|
|             Olly W.|seller s delivery...|black 50pcs|2023-05-18|         5|
|           Evelyn W.|good quality pric...|black 50pcs|2023-05-09|         5|
|           Connie L.|but really differ...|black 50pcs|2024-01-20|         2|
|              Woo J.|             no good|black 50pcs|2022-07-18|         2|
|               V***.|fast shipment ite...|black 50pcs|2023-07-16|         5|
|               I***.|              thanks|black 50pcs|2023-09-22|         5|
|            Fenhong.|quality very sati...|black 50pcs|2023-04-25|         5|
|          Jesmond K.|stew oriya uses g...|black 50pcs|2023-01-15|         5|
|               H***.|despite seller ta...|black 50pcs|2022-08-30|         5|
|              For B.|good quality reas...|black 50pcs|2022-08-31|         5|
|            AZRIN B.|received goods co...|black 50pcs|2022-09-10|         5|
|               HS L.|    thank you seller|black 50pcs|2022-10-18|         5|
|               C***.|thank you face ma...|black 50pcs|2024-01-17|         3|
|               1***9|                good|black 50pcs|2023-10-06|         5|
|              saiful|         goods until|black 50pcs|2023-01-13|         5|
|               0***6|received goods go...|black 50pcs|2022-11-23|         5|
|                Emma|fast response rec...|black 50pcs|2023-10-18|         5|
|           Celine L.|        worth buying|black 50pcs|2022-11-09|         5|
|               T***.|received good con...|black 50pcs|2022-10-15|         5|
|     Geethanjalli S.|        good product|black 50pcs|2022-07-29|         5|
|          Theresa T.|thanks seller sur...|black 50pcs|2023-03-16|         5|
|               b***.|only protection n...|black 50pcs|2022-07-23|         2|
|               C***.|good product good...|black 50pcs|2022-07-18|         5|
|              Ann K.|i already receive...|black 50pcs|2023-12-14|         5|
|            nahar J.|i bought pcs mask...|black 50pcs|2022-09-07|         5|
|             Nick C.|boss goods have a...|black 50pcs|2022-07-15|         5|
|            Karen S.|        poor quality|black 50pcs|2022-08-11|         1|
|                 Saf|super love mask f...|black 50pcs|2022-12-14|        5|
|              Nur I.|        good product|black 50pcs|2024-01-02|        5|
|               L***.|goods are receive...|black 50pcs|2024-04-20|        5|
|            Brian W.|budget mask overa...|black 50pcs|2022-12-12|        5|
|            Sudha C.|good quality comp...|black 50pcs|2022-07-15|        5|
|          Salimah H.|thank you mask su...|black 50pcs|2022-08-15|        5|
|               P***.|fast delivery ite...|black 50pcs|2023-02-22|        5|
|              Umi B.|                nice|black 50pcs|2022-08-15|        5|
|             Wong S.| goods well received|black 50pcs|2022-10-01|        5|
|         Thewaran M.|my time order sel...|black 50pcs|2022-09-21|        5|
|               F***.|received good con...|black 50pcs|2022-10-15|        5|
|          Kenneth C.|fast delivery alw...|black 50pcs|2022-06-29|        5|
|            Leong C.|           thin mask|black 50pcs|2022-07-29|        3|
|          Zayaana S.|really quick deli...|black 50pcs|2022-07-25|        5|
|             Maha D.|received within days|black 50pcs|2023-10-02|        5|
|               P***.|poor quality thin...|black 50pcs|2022-10-18|        2|
|              Tan K.|           beautiful|black 50pcs|2023-09-02|        5|
|            Jason K.|                good|black 50pcs|2022-07-13|        5|
|             Siow H.|comfortable not t...|black 50pcs|2023-08-23|        5|
|               J***.|very good seller ...|black 50pcs|2023-01-07|        5|
|              Lam K.|good price ok too...|black 50pcs|2023-08-16|        5|
+--------------------+--------------------+-----------+----------+---------+
only showing top 100 rows

Connection to Neo4j closed.
```

In [35]:
```
df.count()
```

Out[35]:
```
2139
```

Data Annotation

In [36]:
```
sc.addFile("/home/student/G3_B/de_classes/data_preparation/data_annotation.py")
```

In [37]:
```
# Import necessary libraries
from data_annotation import DataAnnotator  # Adjust the import path if needed

annotator = DataAnnotator(spark_session=spark)
```

```
# Apply the annotation
df_annotated = annotator.add_sentiment_column(df)
df_annotated.show()
```

```
+----------+--------------------+-----------+----------+---------+---------+
|      Name|              Review|    SkuInfo|      Date|StarCount|Sentiment|
+----------+--------------------+-----------+----------+---------+---------+
|     Nasih|             awesome|black 50pcs|2022-07-27|        5|        2|
|Abinash M.|                  ok|black 50pcs|2022-09-28|        5|        2|
|    Loh W.|   great design love|black 50pcs|2024-06-10|        5|        2|
|     Md A.|        good product|black 50pcs|2023-07-15|        5|        2|
| Garlic M.|doesnt match vide...|black 50pcs|2022-07-25|        1|        0|
|  NorHa S.|black mask there ...|black 50pcs|2022-08-16|        5|        2|
|     n***i|received good con...|black 50pcs|2022-06-28|        5|        2|
|THERESA H.|fast delivery but...|black 50pcs|2022-10-03|        5|        2|
|*******896|short three packs...|black 50pcs|2023-05-12|        4|        2|
|  Chris C.|got packing no so...|black 50pcs|2023-09-26|        5|        2|
|    Yan E.|goods have been r...|black 50pcs|2023-08-29|        5|        2|
|    Greedy|           very thin|black 50pcs|2022-11-27|        5|        2|
|     1***1|i ordered sept ch...|black 50pcs|2022-11-14|        5|        2|
|     T***.|good service fast...|black 50pcs|2024-03-03|        5|        2|
|    AgnesQ|inner layer mask ...|black 50pcs|2023-02-27|        5|        2|
|    tan F.|no good quality l...|black 50pcs|2022-07-29|        1|        0|
|   Arse A.|congratulations u...|black 50pcs|2022-07-06|        5|        2|
|     L***.|nice mask fast de...|black 50pcs|2022-07-23|        5|        2|
|     GO A.|not ply made chin...|black 50pcs|2023-02-25|        1|        0|
|   Tang L.|everything good f...|black 50pcs|2023-06-09|        5|        2|
+----------+--------------------+-----------+----------+---------+---------+
only showing top 20 rows
```

In [38]:

```
total_records = df_annotated.count()
print(f"Total number of records: {total_records}")
```

```
Total number of records: 2139
```

In [39]:

```
# Calculate the count of each sentiment category
df_sentiment_distribution = df_annotated.groupBy('Sentiment').count()
df_sentiment_distribution.show()
```

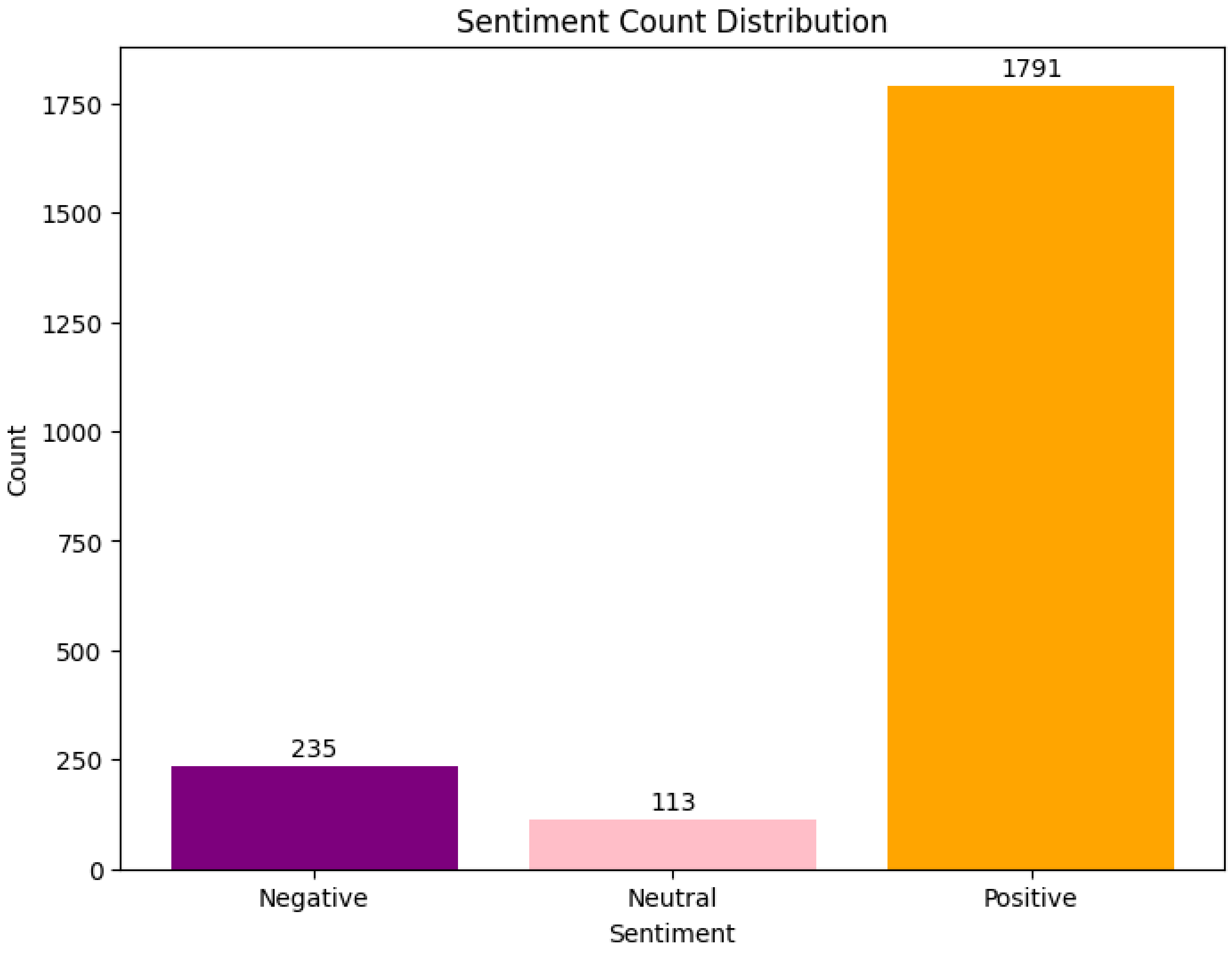
```
+---------+-----+
|Sentiment|count|
+---------+-----+
|        1|  113|
|        2| 1791|
|        0|  235|
+---------+-----+
```

In [40]:

```
sc.addFile("../de_classes/data_visualisation/sentimentVisual.py")
```

In [41]:

```
from sentimentVisual import SentimentPlotter
plotter_count = SentimentPlotter(df_sentiment_distribution, data_type='count')
sentiments, counts = plotter_count.extract_data()
plotter_count.plot_data(sentiments, counts)
```



In [42]:
```
df_annonated.createOrReplaceTempView("labelled_data")
sqlDF = spark.sql("SELECT * FROM labelled_data")
sqlDF.show()
```

```
+----------+--------------------+-----------+----------+---------+---------+
|      Name|              Review|    SkuInfo|      Date|StarCount|Sentiment|
+----------+--------------------+-----------+----------+---------+---------+
|     Nasih|             awesome|black 50pcs|2022-07-27|        5|        2|
|Abinash M.|                  ok|black 50pcs|2022-09-28|        5|        2|
|    Loh W.|   great design love|black 50pcs|2024-06-10|        5|        2|
|     Md A.|        good product|black 50pcs|2023-07-15|        5|        2|
| Garlic M.|doesnt match vide...|black 50pcs|2022-07-25|        1|        0|
|  NorHa S.|black mask there ...|black 50pcs|2022-08-16|        5|        2|
|     n***i|received good con...|black 50pcs|2022-06-28|        5|        2|
|THERESA H.|fast delivery but...|black 50pcs|2022-10-03|        5|        2|
|*******896|short three packs...|black 50pcs|2023-05-12|        4|        2|
|  Chris C.|got packing no so...|black 50pcs|2023-09-26|        5|        2|
|    Yan E.|goods have been r...|black 50pcs|2023-08-29|        5|        2|
|    Greedy|           very thin|black 50pcs|2022-11-27|        5|        2|
|     1***1|i ordered sept ch...|black 50pcs|2022-11-14|        5|        2|
|     T***.|good service fast...|black 50pcs|2024-03-03|        5|        2|
|    AgnesQ|inner layer mask ...|black 50pcs|2023-02-27|        5|        2|
|    tan F.|no good quality l...|black 50pcs|2022-07-29|        1|        0|
|   Arse A.|congratulations u...|black 50pcs|2022-07-06|        5|        2|
|     L***.|nice mask fast de...|black 50pcs|2022-07-23|        5|        2|
|     GO A.|not ply made chin...|black 50pcs|2023-02-25|        1|        0|
|   Tang L.|everything good f...|black 50pcs|2023-06-09|        5|        2|
+----------+--------------------+-----------+----------+---------+---------+
only showing top 20 rows
```

Positive and Negative Dataframe, Filter Out Neutral Records

In [43]:

```
no_neutral_df = spark.sql("SELECT * FROM labelled_data WHERE Sentiment != 1")
no_neutral_df.show()
```

```
+----------+--------------------+-----------+----------+---------+---------+
|      Name|              Review|    SkuInfo|      Date|StarCount|Sentiment|
+----------+--------------------+-----------+----------+---------+---------+
|     Nasih|             awesome|black 50pcs|2022-07-27|        5|        2|
|Abinash M.|                  ok|black 50pcs|2022-09-28|        5|        2|
|    Loh W.|   great design love|black 50pcs|2024-06-10|        5|        2|
|     Md A.|        good product|black 50pcs|2023-07-15|        5|        2|
| Garlic M.|doesnt match vide...|black 50pcs|2022-07-25|        1|        0|
|  NorHa S.|black mask there ...|black 50pcs|2022-08-16|        5|        2|
|     n***i|received good con...|black 50pcs|2022-06-28|        5|        2|
|THERESA H.|fast delivery but...|black 50pcs|2022-10-03|        5|        2|
|*******896|short three packs...|black 50pcs|2023-05-12|        4|        2|
|  Chris C.|got packing no so...|black 50pcs|2023-09-26|        5|        2|
|    Yan E.|goods have been r...|black 50pcs|2023-08-29|        5|        2|
|    Greedy|           very thin|black 50pcs|2022-11-27|        5|        2|
|     1***1|i ordered sept ch...|black 50pcs|2022-11-14|        5|        2|
|     T***.|good service fast...|black 50pcs|2024-03-03|        5|        2|
|    AgnesQ|inner layer mask ...|black 50pcs|2023-02-27|        5|        2|
|    tan F.|no good quality l...|black 50pcs|2022-07-29|        1|        0|
|   Arse A.|congratulations u...|black 50pcs|2022-07-06|        5|        2|
|     L***.|nice mask fast de...|black 50pcs|2022-07-23|        5|        2|
|     GO A.|not ply made chin...|black 50pcs|2023-02-25|        1|        0|
|   Tang L.|everything good f...|black 50pcs|2023-06-09|        5|        2|
+----------+--------------------+-----------+----------+---------+---------+
only showing top 20 rows
```

In [44]:
```
total_records = no_neutral_df.count()
print(f"Total number of records: {total_records}")
```

```
Total number of records: 2026
```

In [45]:
```
#Calculate the count of each sentiment category
df_sentiment_distribution_new = no_neutral_df.groupBy('Sentiment').count()
df_sentiment_distribution_new.show()
```

```
+---------+-----+
|Sentiment|count|
+---------+-----+
|        2| 1791|
|        0|  235|
+---------+-----+
```

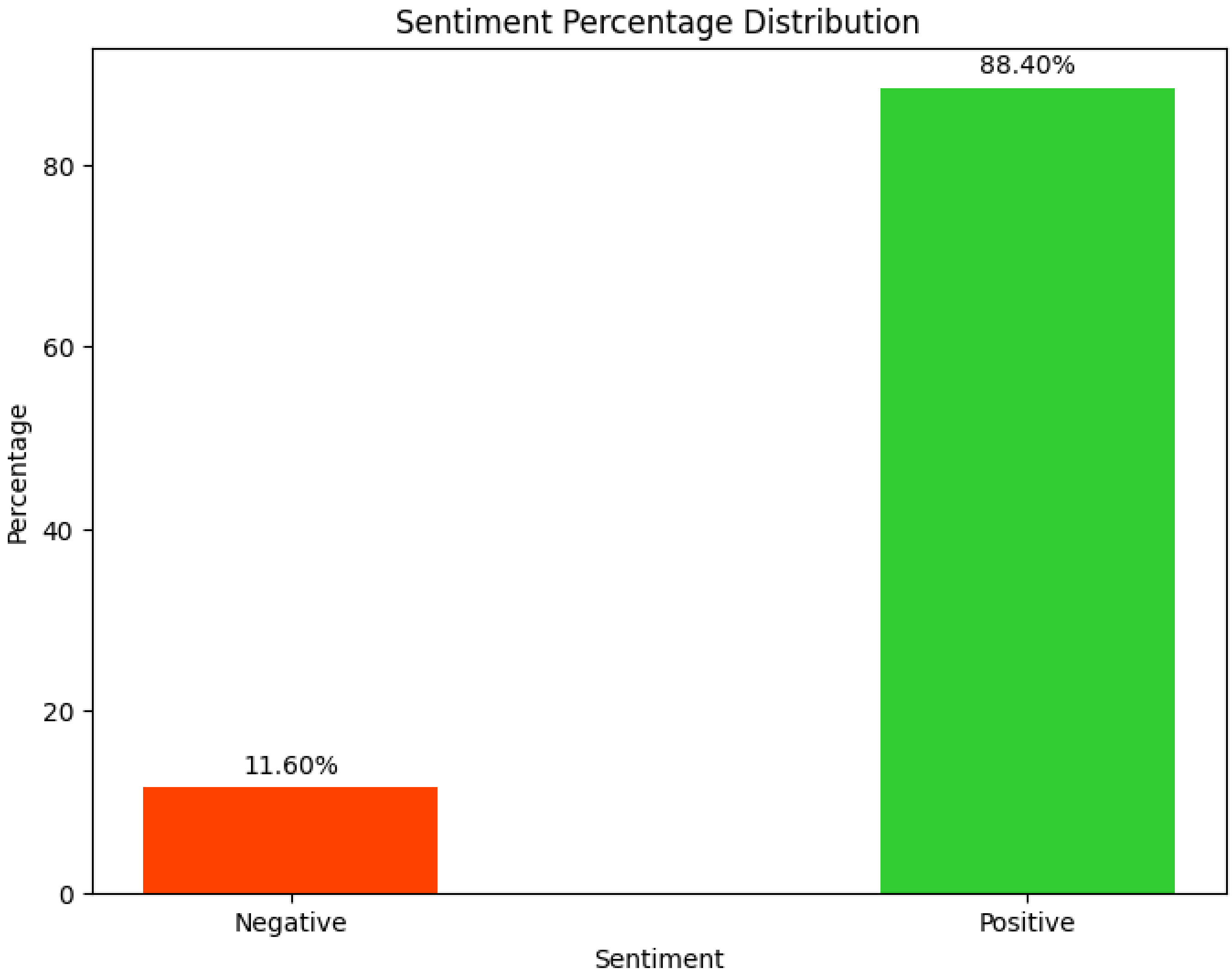
In [46]:

```python
from pyspark.sql.functions import col

total_count = no_neutral_df.count()

# Calculate the percentage for each sentiment category
df_sentiment_distribution_new = df_sentiment_distribution_new.withColumn(
    "Percentage",
    (col("count") / total_count) * 100
)
df_sentiment_distribution_new.show()
```

```
+---------+-----+------------------+
|Sentiment|count|        Percentage|
+---------+-----+------------------+
|        2| 1791| 88.40078973346496|
|        0|  235|11.599210266535044|
+---------+-----+------------------+
```

In [47]:

```python
from sentimentVisual import SentimentPlotter
plotter_percentage = SentimentPlotter(df_sentiment_distribution_new, data_type='percentage')
sentiments, percentages = plotter_percentage.extract_data()
plotter_percentage.plot_data(sentiments, percentages)
```



MongoDB

In [48]:
```
sc.addFile("/home/student/G3_B/de_classes/data_storage/mongodb_handler.py")
```

In [49]:

```
from mongodb_handler import MongoDBHandler

# Initialize the MongoDB handler
mongo_handler = MongoDBHandler(
    uri="mongodb+srv://mongo:456456@cluster0.muusd.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0",
    database="123",
    collection="reviews"
)
```

Store (Skip)

Please skip this step as we have already save it into mongo. Please run the code for data load

In [50]:

```
# #Convert PySpark DataFrame to List of Dictionaries
# data_dict = no_neutral_df.toPandas().to_dict("records")
```

In [51]:

```
# from datetime import datetime, date

# #Convert the 'Date' field to string format for each dictionary in the list
# for record in data_dict:
#     if isinstance(record['Date'], date):  # 'date' refers to the datetime.date class
#         record['Date'] = record['Date'].strftime('%Y-%m-%d')
```

In [52]:

```
# collection.insert_many(data_dict)

# print("Data inserted successfully into MongoDB!")
```

In [53]:

```
# List some documents
mongo_handler.list_documents(limit=3)
# Retrieve data from MongoDB
mongo_data = mongo_handler.retrieve_data()
```

```
Listing 3 documents in the reviews collection:
[{'Date': '2023-08-29',
  'Name': 'Yan E.',
  'Review': 'goods have been received thank you seller',
  'Sentiment': 2,
  'SkuInfo': 'black 50pcs',
  'StarCount': 5,
  '_id': ObjectId('66d406b441c47d9ca376c1f4')},
 {'Date': '2022-05-30',
  'Name': 'Thevagi S.',
  'Review': 'all good',
  'Sentiment': 2,
  'SkuInfo': 'grey',
  'StarCount': 5,
  '_id': ObjectId('66d406b441c47d9ca376c1f5')},
 {'Date': '2022-04-03',
  'Name': 'BC K.',
  'Review': 'products packaging not see through can t see mask colour inside',
  'Sentiment': 2,
  'SkuInfo': 'black',
  'StarCount': 4,
  '_id': ObjectId('66d406b441c47d9ca376c1f6')}]
```

Load from MongoDB to Dataframe

In [54]:

```python
# Define the schema for the PySpark DataFrame
from pyspark.sql.types import StructType, StructField, StringType, IntegerType

schema = StructType([
    StructField("Name", StringType(), True),
    StructField("Review", StringType(), True),
    StructField("SkuInfo", StringType(), True),
    StructField("Date", StringType(), True),
    StructField("StarCount", IntegerType(), True),
    StructField("Sentiment", IntegerType(), True)
])

df_loaded_mongo = mongo_handler.convert_to_dataframe(mongo_data, spark, schema)
df_loaded_mongo.show()
```

```
+----------+--------------------+----------------+----------+---------+---------+
|      Name|              Review|         SkuInfo|      Date|StarCount|Sentiment|
+----------+--------------------+----------------+----------+---------+---------+
|    Yan E.|goods have been r...|     black 50pcs|2023-08-29|        5|        2|
|Thevagi S.|            all good|            grey|2022-05-30|        5|        2|
|     BC K.|products packagin...|           black|2022-04-03|        4|        2|
| Jennie 8.|good quality prod...|           black|2022-08-01|        5|        2|
|     S***.|touch hold clippe...|           black|2024-01-29|        5|        2|
|  Eddie T.|good seller fast ...|careion 3d black|2023-09-18|        5|        2|
|    Thi N.|i think good me i...|           white|2023-08-31|        5|        2|
|  Narainis|likes love sorry ...|           black|2023-04-05|        5|        2|
|   Khor S.|     thanks received|           black|2022-12-21|        5|        2|
|    Greedy|           very thin|     black 50pcs|2022-11-27|        5|        2|
|    Yap B.|fast shipping wil...|           black|2023-06-05|        5|        2|
|     W***.|happy thanks sell...|           black|2022-05-28|        5|        2|
|     N***.|mask enough messa...|            grey|2022-04-01|        5|        2|
|     K***.|strange see anoth...| headloop purple|2022-04-10|        5|        2|
|Theresa T.|item just receive...|     white 50pcs|2023-11-26|        5|        2|
|     1***1|i ordered sept ch...|     black 50pcs|2022-11-14|        5|        2|
|    Lun K.|first time buy fe...|           white|2022-04-25|        5|        2|
|     L***.|mask big enough m...|      grey 50pcs|2022-09-12|        5|        2|
|  Ahmad C.|received good ord...|     white 50pcs|2022-09-05|        5|        2|
|     So P.|price cheaper but...|           white|2022-05-13|        4|        2|
+----------+--------------------+----------------+----------+---------+---------+
only showing top 20 rows
```

In [55]:

```
total_rows = df_loaded_mongo.count()
print(f"Total number of rows: {total_rows}")
```

```
Total number of rows: 2026
```

In [56]:

```
mongo_handler.close()
```

One Hot Encoding

In [57]:

```
sc.addFile("../de_classes/data_preparation/data_transformation.py")
```

In [58]:
```
from data_transformation import DataTransformations
transform = DataTransformations()
```

In [59]:
```
df_encoded = DataTransformations.one_hot_encode(df_loaded_mongo, 'SkuInfo')
```

In [60]:
```
df_encoded.show()
```

```
+----------+--------------------+----------------+----------+---------+---------+-------------+---------------+
|      Name|              Review|         SkuInfo|      Date|StarCount|Sentiment|SkuInfo_index|SkuInfo_encoded|
+----------+--------------------+----------------+----------+---------+---------+-------------+---------------+
|    Yan E.|goods have been r...|     black 50pcs|2023-08-29|        5|        2|          2.0| (53,[2],[1.0])|
|Thevagi S.|            all good|            grey|2022-05-30|        5|        2|          4.0| (53,[4],[1.0])|
|     BC K.|products packagin...|           black|2022-04-03|        4|        2|          0.0| (53,[0],[1.0])|
| Jennie 8.|good quality prod...|           black|2022-08-01|        5|        2|          0.0| (53,[0],[1.0])|
|     S***.|touch hold clippe...|           black|2024-01-29|        5|        2|          0.0| (53,[0],[1.0])|
|  Eddie T.|good seller fast ...|careion 3d black|2023-09-18|        5|        2|         13.0|(53,[13],[1.0])|
|    Thi N.|i think good me i...|           white|2023-08-31|        5|        2|          1.0| (53,[1],[1.0])|
|  Narainis|likes love sorry ...|           black|2023-04-05|        5|        2|          0.0| (53,[0],[1.0])|
|   Khor S.|     thanks received|           black|2022-12-21|        5|        2|          0.0| (53,[0],[1.0])|
|    Greedy|           very thin|     black 50pcs|2022-11-27|        5|        2|          2.0| (53,[2],[1.0])|
|    Yap B.|fast shipping wil...|           black|2023-06-05|        5|        2|          0.0| (53,[0],[1.0])|
|     W***.|happy thanks sell...|           black|2022-05-28|        5|        2|          0.0| (53,[0],[1.0])|
|     N***.|mask enough messa...|            grey|2022-04-01|        5|        2|          4.0| (53,[4],[1.0])|
|     K***.|strange see anoth...| headloop purple|2022-04-10|        5|        2|         26.0|(53,[26],[1.0])|
|Theresa T.|item just receive...|     white 50pcs|2023-11-26|        5|        2|          3.0| (53,[3],[1.0])|
|     1***1|i ordered sept ch...|     black 50pcs|2022-11-14|        5|        2|          2.0| (53,[2],[1.0])|
|    Lun K.|first time buy fe...|           white|2022-04-25|        5|        2|          1.0| (53,[1],[1.0])|
|     L***.|mask big enough m...|      grey 50pcs|2022-09-12|        5|        2|          8.0| (53,[8],[1.0])|
|  Ahmad C.|received good ord...|     white 50pcs|2022-09-05|        5|        2|          3.0| (53,[3],[1.0])|
|     So P.|price cheaper but...|           white|2022-05-13|        4|        2|          1.0| (53,[1],[1.0])|
+----------+--------------------+----------------+----------+---------+---------+-------------+---------------+
only showing top 20 rows
```

In [61]:
```
from pyspark.ml.feature import StringIndexer
df_encoded.select("SkuInfo", "SkuInfo_index").show(truncate=False)
```

```
+----------------+-------------+
|SkuInfo         |SkuInfo_index|
+----------------+-------------+
|black 50pcs     |2.0          |
|grey            |4.0          |
|black           |0.0          |
|black           |0.0          |
|black           |0.0          |
|careion 3d black|13.0         |
|white           |1.0          |
|black           |0.0          |
|black           |0.0          |
|black 50pcs     |2.0          |
|black           |0.0          |
|black           |0.0          |
|grey            |4.0          |
|headloop purple |26.0         |
|white 50pcs     |3.0          |
|black 50pcs     |2.0          |
|white           |1.0          |
|grey 50pcs      |8.0          |
|white 50pcs     |3.0          |
|white           |1.0          |
+----------------+-------------+
only showing top 20 rows
```

Tokenization

In [62]:

```
df_tokenized = DataTransformations.tokenize(df_encoded, input_col="Review", output_col="tokens")
```

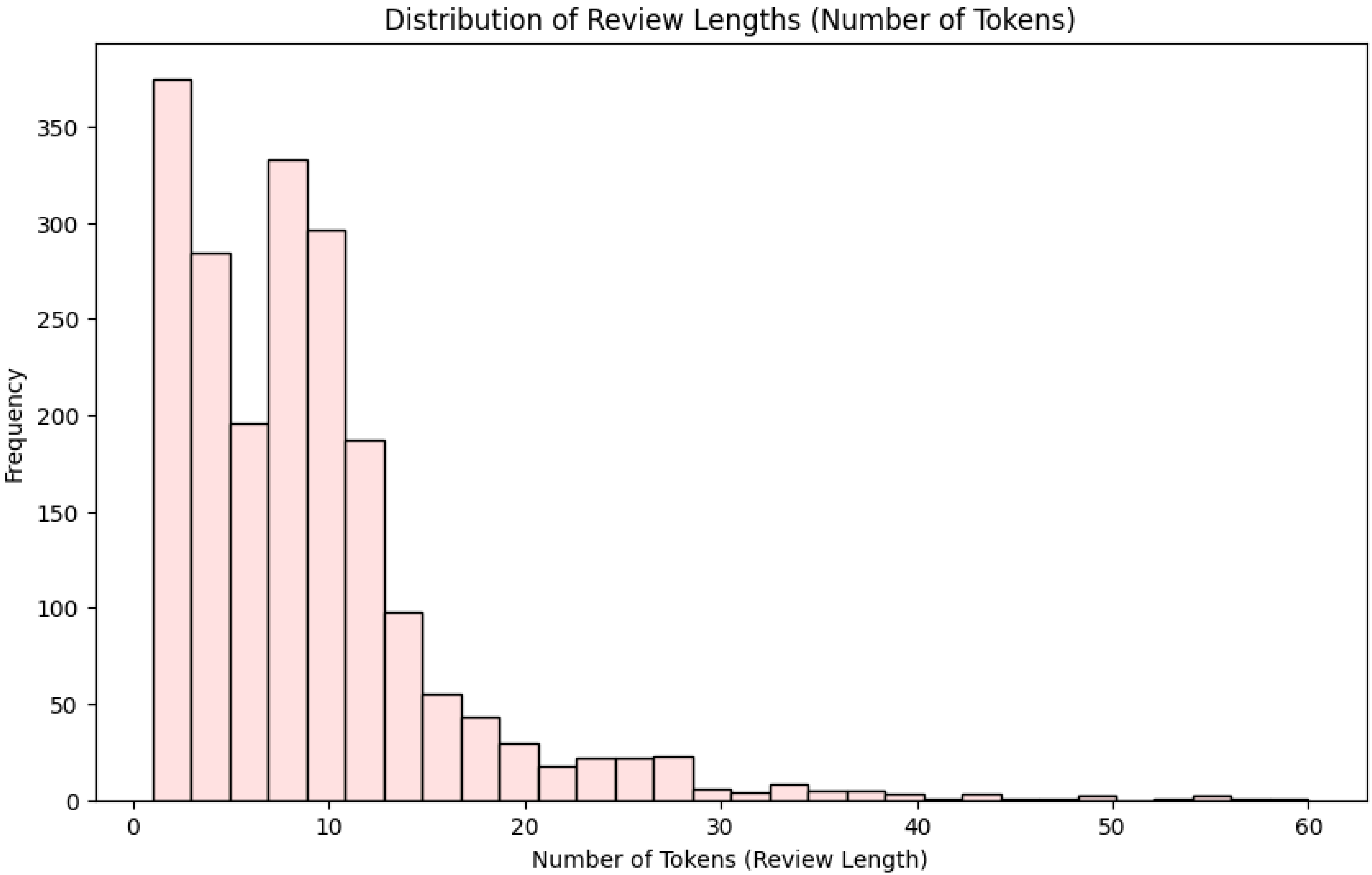
In [63]:

```
df_tokenized.select("Review", "tokens").show(truncate=True)
```

```
+--------------------+--------------------+
|              Review|              tokens|
+--------------------+--------------------+
|goods have been r...|[goods, have, bee...|
|            all good|         [all, good]|
|products packagin...|[products, packag...|
|good quality prod...|[good, quality, p...|
|touch hold clippe...|[touch, hold, cli...|
|good seller fast ...|[good, seller, fa...|
|i think good me i...|[i, think, good, ...|
|likes love sorry ...|[likes, love, sor...|
|     thanks received|  [thanks, received]|
|           very thin|        [very, thin]|
|fast shipping wil...|[fast, shipping, ...|
|happy thanks sell...|[happy, thanks, s...|
|mask enough messa...|[mask, enough, me...|
|strange see anoth...|[strange, see, an...|
|item just receive...|[item, just, rece...|
|i ordered sept ch...|[i, ordered, sept...|
|first time buy fe...|[first, time, buy...|
|mask big enough m...|[mask, big, enoug...|
|received good ord...|[received, good, ...|
|price cheaper but...|[price, cheaper, ...|
+--------------------+--------------------+
only showing top 20 rows
```

In [64]:
```
sc.addFile("../de_classes/data_visualisation/reviewLength.py")
```

In [65]:
```
from reviewLength import ReviewLengthAnalyzer

analyzer = ReviewLengthAnalyzer(df_tokenized)
df_review_length = analyzer.calculate_review_lengths()

review_lengths = analyzer.collect_review_lengths(df_review_length)
analyzer.plot_review_length_distribution(review_lengths)
```



Lemmatization

In [66]:
```
import nltk
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to /home/student/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[66]:

```
True
```

In [67]:

```
df_lemmatized = DataTransformations.lemmatize_tokens(df_tokenized, tokens_col="tokens")
```

In [68]:

```
df_lemmatized.show(truncate=True)
```

```
[Stage 147:>                                                        (0 + 1) / 1]
```

```
+----------+--------------------+----------------+----------+---------+---------+-------------+---------------+---------------
----+
|      Name|              Review|         SkuInfo|      Date|StarCount|Sentiment|SkuInfo_index|SkuInfo_encoded|              to
kens|
+----------+--------------------+----------------+----------+---------+---------+-------------+---------------+---------------
----+
|    Yan E.|goods have been r...|     black 50pcs|2023-08-29|        5|        2|          2.0| (53,[2],[1.0])|[good, have, bee
n...|
|Thevagi S.|            all good|            grey|2022-05-30|        5|        2|          4.0| (53,[4],[1.0])|         [all, g
ood]|
|     BC K.|products packagin...|           black|2022-04-03|        4|        2|          0.0| (53,[0],[1.0])|[product, packag
i...|
| Jennie 8.|good quality prod...|           black|2022-08-01|        5|        2|          0.0| (53,[0],[1.0])|[good, quality, 
p...|
|     S***.|touch hold clippe...|           black|2024-01-29|        5|        2|          0.0| (53,[0],[1.0])|[touch, hold, cl
i...|
|  Eddie T.|good seller fast ...|careion 3d black|2023-09-18|        5|        2|         13.0|(53,[13],[1.0])|[good, seller, f
a...|
|    Thi N.|i think good me i...|           white|2023-08-31|        5|        2|          1.0| (53,[1],[1.0])|[i, think, good,
...|
|  Narainis|likes love sorry ...|           black|2023-04-05|        5|        2|          0.0| (53,[0],[1.0])|[like, love, sor
r...|
|   Khor S.|     thanks received|           black|2022-12-21|        5|        2|          0.0| (53,[0],[1.0])|  [thanks, recei
ved]|
|    Greedy|           very thin|     black 50pcs|2022-11-27|        5|        2|          2.0| (53,[2],[1.0])|        [very, t
hin]|
|    Yap B.|fast shipping wil...|           black|2023-06-05|        5|        2|          0.0| (53,[0],[1.0])|[fast, shipping,
...|
|     W***.|happy thanks sell...|           black|2022-05-28|        5|        2|          0.0| (53,[0],[1.0])|[happy, thanks, 
s...|
|     N***.|mask enough messa...|            grey|2022-04-01|        5|        2|          4.0| (53,[4],[1.0])|[mask, enough, m
e...|
|     K***.|strange see anoth...| headloop purple|2022-04-10|        5|        2|         26.0|(53,[26],[1.0])|[strange, see, a
n...|
|Theresa T.|item just receive...|     white 50pcs|2023-11-26|        5|        2|          3.0| (53,[3],[1.0])|[item, just, rec
e...|
|     1***1|i ordered sept ch...|     black 50pcs|2022-11-14|        5|        2|          2.0| (53,[2],[1.0])|[i, ordered, sep
t...|
|    Lun K.|first time buy fe...|           white|2022-04-25|        5|        2|          1.0| (53,[1],[1.0])|[first, time, bu
y...|
|     L***.|mask big enough m...|      grey 50pcs|2022-09-12|        5|        2|          8.0| (53,[8],[1.0])|[mask, big, enou
```

```
g...|
|  Ahmad C.|received good ord...|    white 50pcs|2022-09-05|        5|        2|          3.0| (53,[3],[1.0])|[received, good,
...|
|     So P.|price cheaper but...|           white|2022-05-13|        4|        2|          1.0| (53,[1],[1.0])|[price, cheaper,
...|
+----------+--------------------+----------------+----------+---------+---------+-------------+---------------+----------------
----+
only showing top 20 rows
```

DROP irrelevant columns

In [69]:
```python
df_relevant_columns = df_lemmatized.drop('Name','SkuInfo','Date','StarCount','SkuInfo_encoded')
```

In [70]:
```python
from pyspark.sql.functions import size
# Create a new column "number_of_tokens" by calculating the length of the "tokens" array
df_with_token_count = df_relevant_columns.withColumn("number_of_tokens", size("tokens"))
```

In [71]:
```python
df_with_token_count.select("Review","tokens","number_of_tokens").show(truncate=False)
```

```
[Stage 148:>                                                        (0 + 1) / 1]
```

```
+-------------------------------------------------------------------------------------------------------------------------------
--------------------+-----------------------------------------------------------------------------------------------------------
------------------------------------------------------------------+----------------+
|Review
|tokens
|number_of_tokens|
+-------------------------------------------------------------------------------------------------------------------------------
--------------------+-----------------------------------------------------------------------------------------------------------
------------------------------------------------------------------+----------------+
|goods have been received thank you seller
|[good, have, been, received, thank, you, seller]
|7               |
|all good
|[all, good]
|2               |
|products packaging not see through can t see mask colour inside
|[product, packaging, not, see, through, can, t, see, mask, colour, inside]
|11              |
|good quality product excellent services seller
|[good, quality, product, excellent, service, seller]
|6               |
|touch hold clipped content fix unspecked clipping content will be deleted after hour welcome gboard clipboard all texts you co
py will be saved here|[touch, hold, clipped, content, fix, unspecked, clipping, content, will, be, deleted, after, hour, welcom
e, gboard, clipboard, all, text, you, copy, will, be, saved, here]|24              |
|good seller fast deliverygood seller fast deliverygood seller fast delivery
|[good, seller, fast, deliverygood, seller, fast, deliverygood, seller, fast, delivery]
|10              |
|i think good me i will use review later you guys
|[i, think, good, me, i, will, use, review, later, you, guy]
|11              |
|likes love sorry picture not relevant but product all best
|[like, love, sorry, picture, not, relevant, but, product, all, best]
|10              |
|thanks received
|[thanks, received]
|2               |
|very thin
|[very, thin]
|2               |
|fast shipping will repurchase again thank you
|[fast, shipping, will, repurchase, again, thank, you]
```

```
|7               |
|happy thanks seller just little thinner suitable price
|[happy, thanks, seller, just, little, thinner, suitable, price]
|8               |
|mask enough message cheap prices can buy more
|[mask, enough, message, cheap, price, can, buy, more]
|8               |
|strange see another mask order other masks arrive disappointed even cheap other kinds masks tetapi thin giler saya nkanother m
ask you eh           |[strange, see, another, mask, order, other, mask, arrive, disappointed, even, cheap, other, kind, mask, t
etapi, thin, giler, saya, nkanother, mask, you, eh]               |22              |
|item just received today packed neat mask same per advertised comfortable wear thank you lazada prompt delivery
|[item, just, received, today, packed, neat, mask, same, per, advertised, comfortable, wear, thank, you, lazada, prompt, delive
ry]                                          |17              |
|i ordered sept check size design all were good so i re ordered end up different design nose bridge dissappointed
|[i, ordered, sept, check, size, design, all, were, good, so, i, re, ordered, end, up, different, design, nose, bridge, dissapp
ointed]                                      |20              |
|first time buy feel good
|[first, time, buy, feel, good]
|5               |
|mask big enough man s face comfortable good price too thank you seller courier guy
|[mask, big, enough, man, s, face, comfortable, good, price, too, thank, you, seller, courier, guy]
|15              |
|received good order good supplier
|[received, good, order, good, supplier]
|5               |
|price cheaper but thinner layer
|[price, cheaper, but, thinner, layer]
|5               |
+------------------------------------------------------------------------------------------------------------------------------
---------------------+---------------------------------------------------------------------------------------------------------
------------------------------------------------------------------+----------------+
only showing top 20 rows
```

In [72]:

```
df_with_token_count.printSchema()
```

```
root
 |-- Review: string (nullable = true)
 |-- Sentiment: integer (nullable = true)
 |-- SkuInfo_index: double (nullable = false)
 |-- tokens: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- number_of_tokens: integer (nullable = false)
```

Train Test Split

In [73]:

```
# Split the DataFrame into train and test sets
train_df, test_df = df_with_token_count.randomSplit([0.7, 0.3], seed=30)
```

In [74]:

```
train_df.count()
```

Out[74]:

```
1382
```

In [75]:

```
test_df.count()
```

Out[75]:

```
644
```

Hbase Data Store

In [77]:

```
sc.addFile("/home/student/G3_B/de_classes/data_storage/hbase_handler.py")
from hbase_handler import HBaseHandler
hbase_handler = HBaseHandler(host='localhost', port=9090)# Create an instance
```

```
24/09/07 19:36:34 WARN SparkContext: The path /home/student/G3_B/de_classes/data_storage/hbase_handler.py has been added alread
y. Overwriting of added paths is not supported in the current version.
```

Create Hbase Table

In [81]:
```
# Define the column families
column_families = {
    'cf1': dict(),
    'cf2': dict(),
    'cf3': dict(),
    'cf4': dict(),
    'cf5': dict()
}
# Create the table
hbase_handler.create_table('train_data', column_families)
hbase_handler.create_table('test_data', column_families)
```

```
Table 'train_data' created successfully.
Available tables: [b'train_data']
Table 'test_data' created successfully.
Available tables: [b'test_data', b'train_data']
```

Store In Hbase

In [82]:
```
# Passing the dataframe and the hbase table_name to store the data
hbase_handler.save_to_hbase(train_df, 'train_data')
hbase_handler.save_to_hbase(test_df, 'test_data')
```

```
Data successfully stored in HBase with 1382 records
```

```
Data successfully stored in HBase with 644 records
```

Delete connection table (Optional can skip)

In [80]:
```
#Delete the table (optional) remove # for running the code
#hbase_handler.delete_table('test_data')
```

```
Table 'test_data' disabled successfully.
Table 'test_data' deleted successfully.
Available tables: []
```

Close Hbase Connection

In [83]:

```
# Close the connection
hbase_handler.close()
```

In [84]:

```
spark.stop()
```

In [ ]: