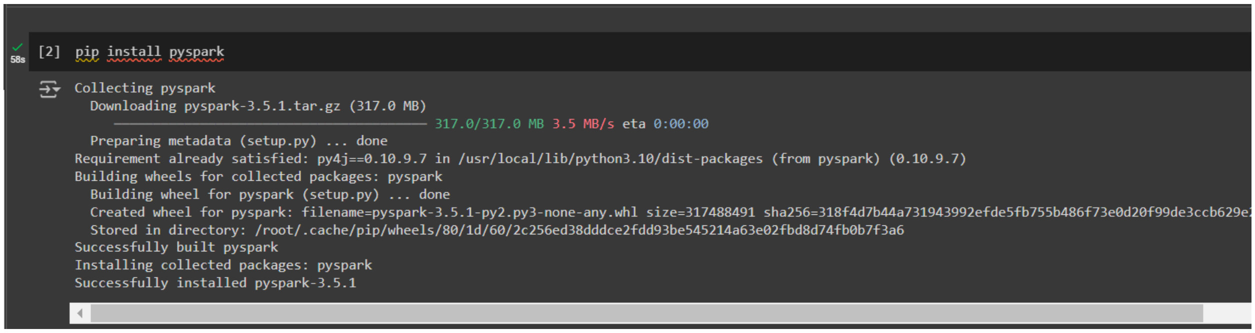
BWT Task-08 Exercise

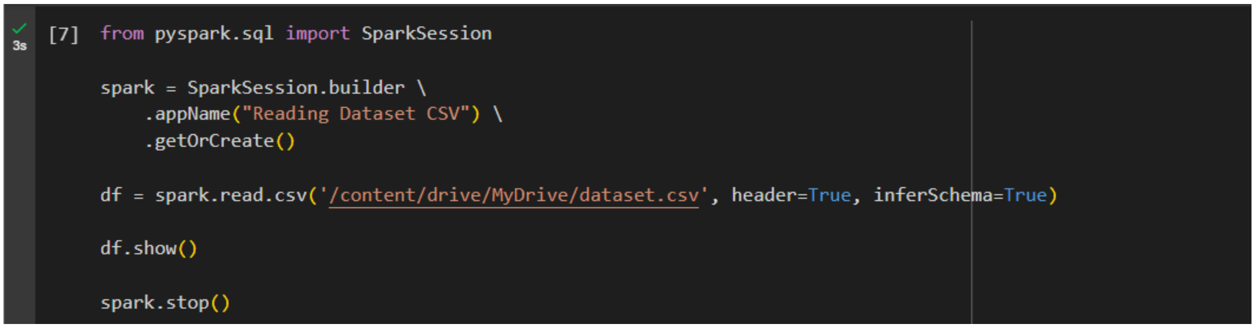
Submitted By: ABRAR SAEED

PySpark:

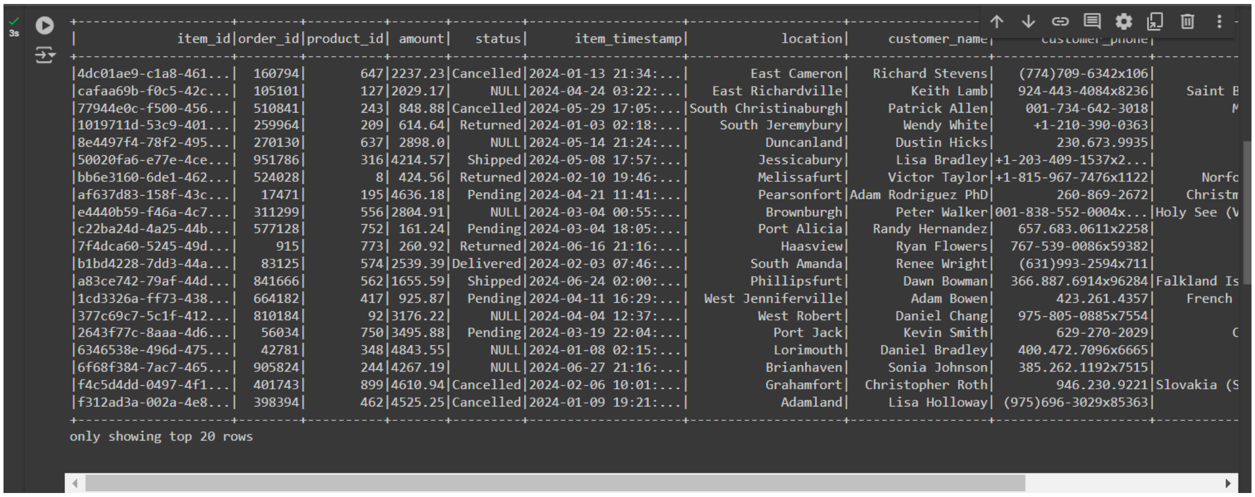
Code & Output:



```
[2] pip install pyspark

Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 3.5 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488491 sha256=318f4d7b44a731943992efde5fb755b486f73e0d20f99de3ccb629e
  Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38dddce2fdd93be545214a63e02fbd8d74fb0b7f3a6
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.1
```

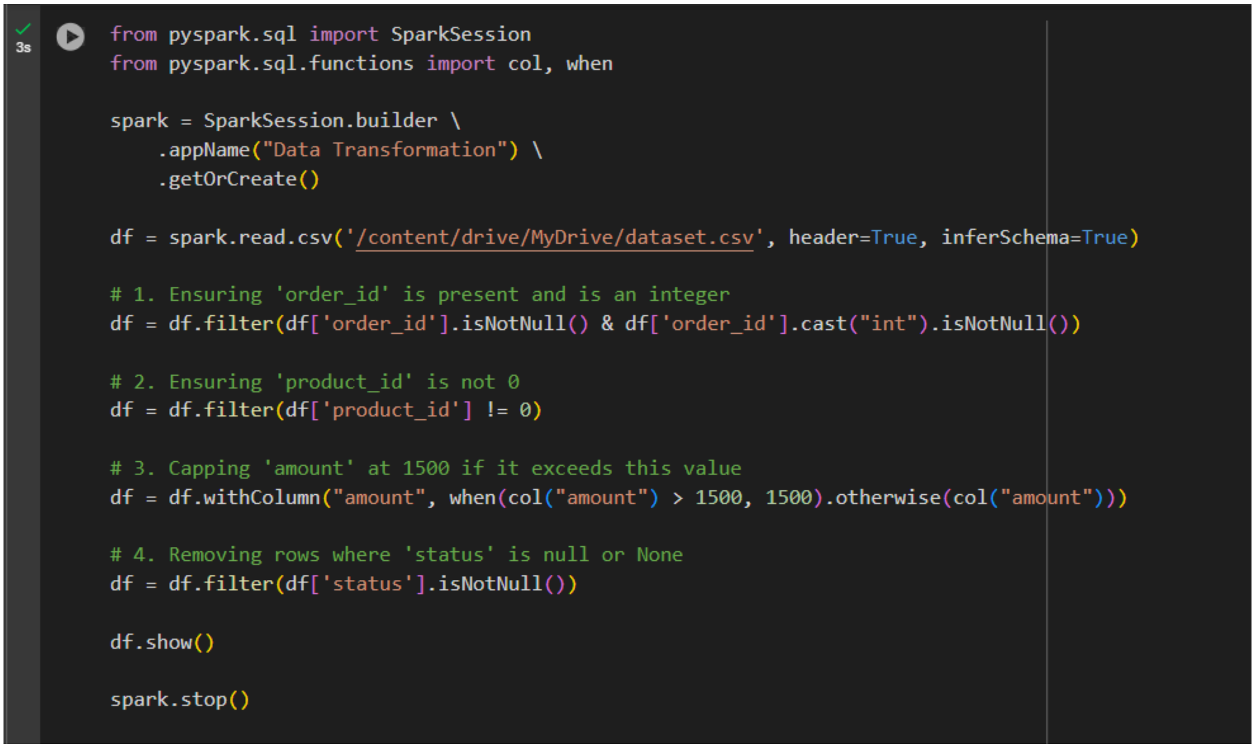


```python
[7] from pyspark.sql import SparkSession

    spark = SparkSession.builder \
        .appName("Reading Dataset CSV") \
        .getOrCreate()

    df = spark.read.csv('/content/drive/MyDrive/dataset.csv', header=True, inferSchema=True)

    df.show()

    spark.stop()
```

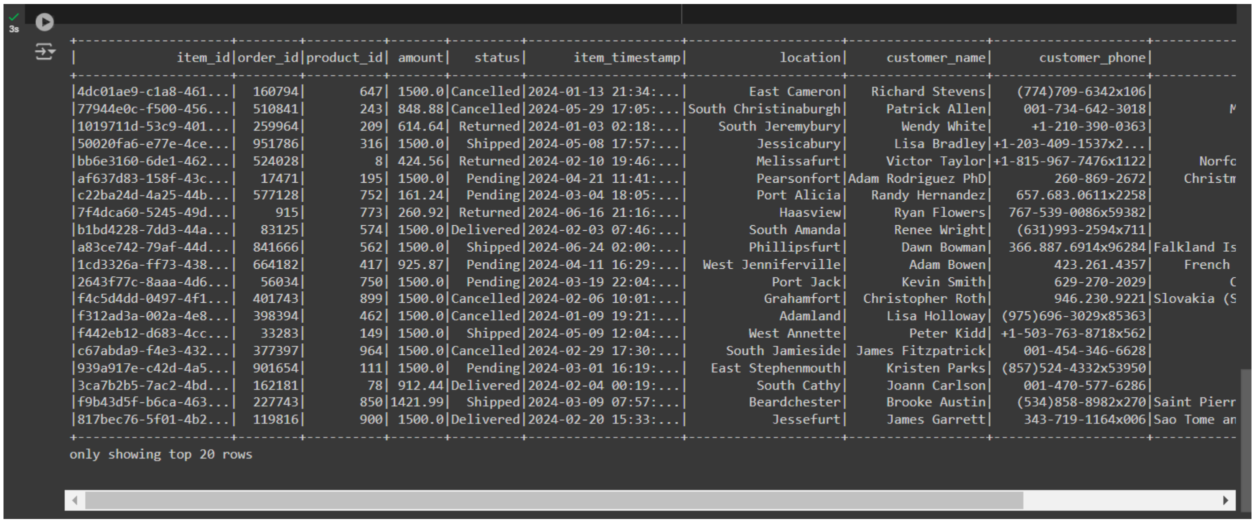


```
+--------------------+--------+----------+-------+---------+-------------------+--------------------+------------------+--------------------+-----------
|             item_id|order_id|product_id| amount|   status|     item_timestamp|            location|     customer_name|      customer_phone|
+--------------------+--------+----------+-------+---------+-------------------+--------------------+------------------+--------------------+-----------
|4dc01ae9-c1a8-461...|  160794|       647|2237.23|Cancelled|2024-01-13 21:34:...|        East Cameron|   Richard Stevens|   (774)709-6342x106|
|cafaa69b-f0c5-42c...|  105101|       127|2029.17|     NULL|2024-04-24 03:22:...|   East Richardville|        Keith Lamb|   924-443-4084x8236|    Saint B
|77944e0c-f500-456...|  510841|       243| 848.88|Cancelled|2024-05-29 17:05:...|South Christinaburgh|     Patrick Allen|    001-734-642-3018|          M
|1019711d-53c9-401...|  259964|       209| 614.64| Returned|2024-01-03 02:18:...|    South Jeremybury|       Wendy White|     +1-210-390-0363|
|8e4497f4-78f2-495...|  270130|       637| 2898.0|     NULL|2024-05-14 21:24:...|          Duncanland|      Dustin Hicks|        230.673.9935|
|50020fa6-e77e-4ce...|  951786|       316|4214.57|  Shipped|2024-05-08 17:57:...|         Jessicabury|      Lisa Bradley|+1-203-409-1537x2...|
|bb6e3160-6de1-462...|  524028|         8| 424.56| Returned|2024-02-10 19:46:...|         Melissafurt|     Victor Taylor|+1-815-967-7476x1122|      Norfo
|af637d83-158f-43c...|   17471|       195|4636.18|  Pending|2024-04-21 11:41:...|         Pearsonfort|Adam Rodriguez PhD|        260-869-2672|    Christm
|e4440b59-f46a-4c7...|  311299|       556|2804.91|     NULL|2024-03-04 00:55:...|          Brownburgh|      Peter Walker|001-838-552-0004x...|Holy See (V
|c22ba24d-4a25-44b...|  577128|       752| 161.24|  Pending|2024-03-04 18:05:...|         Port Alicia|   Randy Hernandez|   657.683.0611x2258|
|7f4dca60-5245-49d...|     915|       773| 260.92| Returned|2024-06-16 21:16:...|            Haasview|      Ryan Flowers|  767-539-0086x59382|
|b1bd4228-7dd3-44a...|   83125|       574|2539.39|Delivered|2024-02-03 07:46:...|        South Amanda|      Renee Wright|   (631)993-2594x711|
|a83ce742-79af-44d...|  841666|       562|1655.59|  Shipped|2024-06-24 02:00:...|        Phillipsfurt|       Dawn Bowman|  366.887.6914x96284|Falkland Is
|1cd3326a-ff73-438...|  664182|       417| 925.87|  Pending|2024-04-11 16:29:...|  West Jenniferville|        Adam Bowen|        423.261.4357|    French
|377c69c7-5c1f-412...|  810184|        92|3176.22|     NULL|2024-04-04 12:37:...|         West Robert|      Daniel Chang|   975-805-0885x7554|
|2643f77c-8aaa-4d6...|   56034|       750|3495.88|  Pending|2024-03-19 22:04:...|           Port Jack|       Kevin Smith|        629-270-2029|          C
|6346538e-496d-475...|   42781|       348|4843.55|     NULL|2024-01-08 02:15:...|           Lorimouth|    Daniel Bradley|   400.472.7096x6665|
|6f68f384-7ac7-465...|  905824|       244|4267.19|     NULL|2024-06-27 21:16:...|          Brianhaven|     Sonia Johnson|   385.262.1192x7515|
|f4c5d4dd-0497-4f1...|  401743|       899|4610.94|Cancelled|2024-02-06 10:01:...|          Grahamfort|  Christopher Roth|        946.230.9221|Slovakia (S
|f312ad3a-002a-4e8...|  398394|       462|4525.25|Cancelled|2024-01-09 19:21:...|            Adamland|     Lisa Holloway| (975)696-3029x85363|
+--------------------+--------+----------+-------+---------+-------------------+--------------------+------------------+--------------------+-----------
only showing top 20 rows
```

The following code does transformations as mentioned in the previous task for this dataset.csv file which were done using pandas. The same thing can be done using PySpark as follows:



```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, when

spark = SparkSession.builder \
    .appName("Data Transformation") \
    .getOrCreate()

df = spark.read.csv('/content/drive/MyDrive/dataset.csv', header=True, inferSchema=True)

# 1. Ensuring 'order_id' is present and is an integer
df = df.filter(df['order_id'].isNotNull() & df['order_id'].cast("int").isNotNull())

# 2. Ensuring 'product_id' is not 0
df = df.filter(df['product_id'] != 0)

# 3. Capping 'amount' at 1500 if it exceeds this value
df = df.withColumn("amount", when(col("amount") > 1500, 1500).otherwise(col("amount")))

# 4. Removing rows where 'status' is null or None
df = df.filter(df['status'].isNotNull())

df.show()

spark.stop()
```



```
+--------------------+--------+----------+-------+---------+--------------------+--------------------+------------------+--------------------+-----------
|             item_id|order_id|product_id| amount|   status|      item_timestamp|            location|     customer_name|      customer_phone|
+--------------------+--------+----------+-------+---------+--------------------+--------------------+------------------+--------------------+-----------
|4dc01ae9-c1a8-461...|  160794|       647| 1500.0|Cancelled|2024-01-13 21:34:...|        East Cameron|   Richard Stevens|   (774)709-6342x106|
|77944e0c-f500-456...|  510841|       243| 848.88|Cancelled|2024-05-29 17:05:...|South Christinaburgh|     Patrick Allen|    001-734-642-3018|          M
|1019711d-53c9-401...|  259964|       209| 614.64| Returned|2024-01-03 02:18:...|    South Jeremybury|       Wendy White|     +1-210-390-0363|
|50020fa6-e77e-4ce...|  951786|       316| 1500.0|  Shipped|2024-05-08 17:57:...|         Jessicabury|      Lisa Bradley|+1-203-409-1537x2...|
|bb6e3160-6de1-462...|  524028|         8| 424.56| Returned|2024-02-10 19:46:...|         Melissafurt|     Victor Taylor|+1-815-967-7476x1122|      Norfo
|af637d83-158f-43c...|   17471|       195| 1500.0|  Pending|2024-04-21 11:41:...|         Pearsonfort|Adam Rodriguez PhD|        260-869-2672|    Christm
|c22ba24d-4a25-44b...|  577128|       752| 161.24|  Pending|2024-03-04 18:05:...|         Port Alicia|   Randy Hernandez|   657.683.0611x2258|
|7f4dca60-5245-49d...|     915|       773| 260.92| Returned|2024-06-16 21:16:...|            Haasview|      Ryan Flowers|  767-539-0086x59382|
|b1bd4228-7dd3-44a...|   83125|       574| 1500.0|Delivered|2024-02-03 07:46:...|        South Amanda|      Renee Wright|   (631)993-2594x711|
|a83ce742-79af-44d...|  841666|       562| 1500.0|  Shipped|2024-06-24 02:00:...|        Phillipsfurt|       Dawn Bowman|  366.887.6914x96284|Falkland Is
|1cd3326a-ff73-438...|  664182|       417| 925.87|  Pending|2024-04-11 16:29:...|  West Jenniferville|        Adam Bowen|        423.261.4357|     French
|2643f77c-8aaa-4d6...|   56034|       750| 1500.0|  Pending|2024-03-19 22:04:...|           Port Jack|       Kevin Smith|        629-270-2029|          C
|f4c5d4dd-0497-4f1...|  401743|       899| 1500.0|Cancelled|2024-02-06 10:01:...|          Grahamfort|  Christopher Roth|        946.230.9221|Slovakia (S
|f312ad3a-002a-4e8...|  398394|       462| 1500.0|Cancelled|2024-01-09 19:21:...|            Adamland|     Lisa Holloway| (975)696-3029x85363|
|f442eb12-d683-4cc...|   33283|       149| 1500.0|  Shipped|2024-05-09 12:04:...|        West Annette|        Peter Kidd| +1-503-763-8718x562|
|c67abda9-f4e3-432...|  377397|       964| 1500.0|Cancelled|2024-02-29 17:30:...|     South Jamieside| James Fitzpatrick|    001-454-346-6628|
|939a917e-c42d-4a5...|  901654|       111| 1500.0|  Pending|2024-03-01 16:19:...|   East Stephenmouth|     Kristen Parks| (857)524-4332x53950|
|3ca7b2b5-7ac2-4bd...|  162181|        78| 912.44|Delivered|2024-02-04 00:19:...|         South Cathy|     Joann Carlson|    001-470-577-6286|
|f9b43d5f-b6ca-463...|  227743|       850|1421.99|  Shipped|2024-03-09 07:57:...|        Beardchester|     Brooke Austin|   (534)858-8982x270|Saint Pierr
|817bec76-5f01-4b2...|  119816|       900| 1500.0|Delivered|2024-02-20 15:33:...|           Jessefurt|     James Garrett|    343-719-1164x006|Sao Tome an
+--------------------+--------+----------+-------+---------+--------------------+--------------------+------------------+--------------------+-----------
only showing top 20 rows
```