NETAJI SUBHAS UNIVERSITY OF TECHNOLOGY

COMPUTER HARDWARE SOFTWARE

WORKSHOP

COCSC19

SUBMITTED BY:

KHUSHI KHARKE- 2021UCS1617

YASH CHAUHAN- 2021UCS1628

ADITYA YADAV- 2021UCS1640

SPARK TASK(UNIT– 4)

Task 1: Explore RDD in spark

RDD stands for Resilient Distributed Dataset and it's one of the fundamental data structures in Apache Spark. RDDs are immutable, distributed collections of objects, which can be processed in parallel across a cluster. They provide fault tolerance, allowing computations to be rerun on failure by using lineage information to rebuild lost data.

Here are some key characteristics and features of RDDs:

1. **Immutability**:
    - RDDs are immutable, meaning once created, their content cannot be changed. This immutability ensures consistency and simplifies parallel processing, as RDDs can be shared across multiple operations without worrying about concurrent modifications.
2. **Distributed Nature**:
    - RDDs are distributed across the nodes of a cluster, allowing for parallel processing of large datasets. This distribution enables Spark to leverage the computing power of multiple machines, making it suitable for handling big data workloads.
3. **Resilience to Failures**:
    - RDDs are resilient to failures through lineage information. Spark keeps track of the sequence of transformations applied to each RDD, allowing it to reconstruct lost partitions by reapplying those transformations. This fault tolerance ensures reliability in distributed computing environments where failures are common.
4. **Lazy Evaluation**:
    - Spark employs lazy evaluation with RDDs, meaning transformations are not executed immediately when called. Instead, Spark builds up a Directed Acyclic Graph (DAG) representing the computation plan. The actual computation occurs only when an action is triggered. Lazy evaluation optimizes performance by allowing Spark to optimize and reorder operations before execution.
5. **In-Memory Computation**:
    - RDDs can be cached or persisted in memory, allowing for faster access during subsequent computations. This in-memory processing capability is crucial for iterative algorithms and interactive data analysis, as it reduces disk I/O overhead and improves overall performance.
6. **Fault Tolerance Mechanism**:

- RDDs achieve fault tolerance through lineage information and resilient distributed storage. When a partition of an RDD is lost due to a node failure, Spark can reconstruct it by reapplying the transformations that led to its creation. This mechanism ensures data integrity and reliability in distributed computations.

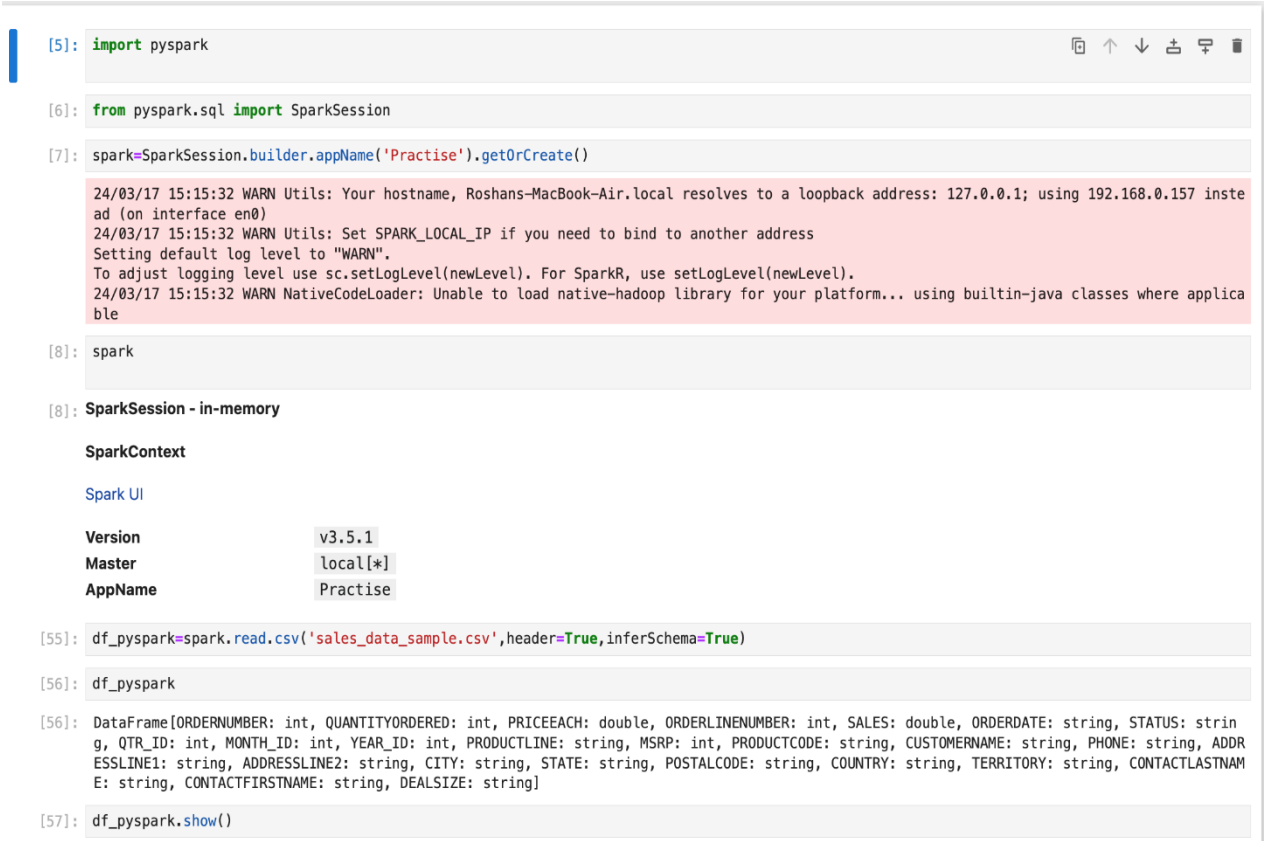
7. **Wide Range of Operations**:
    - RDDs support a wide range of transformations and actions for data manipulation and analysis. Transformations include **map**, **filter**, **flatMap**, **reduceByKey**, **join**, etc., while actions include **collect**, **count**, **reduce**, **foreach**, **saveAsTextFile**, etc. These operations enable complex data processing workflows and analytics tasks.
8. **Persistence Options**:
    - RDDs can be persisted in memory, on disk, or both, using various storage levels (e.g., MEMORY_ONLY, MEMORY_AND_DISK, DISK_ONLY, etc.). Persistence allows RDDs to be reused across multiple computations, reducing redundant computation and improving performance.
9. **Multi-Language Support**:
    - RDDs are accessible from multiple programming languages, including Scala, Java, and Python, making Spark a versatile framework for developers with different language preferences.

Overall, RDDs serve as the foundational building blocks of distributed data processing in Apache Spark, providing resilience, parallelism, and flexibility for handling diverse big data workloads. While DataFrame and Dataset APIs offer higher-level abstractions and optimizations, RDDs remain integral for low-level control and specialized use cases in Spark applications.

Task 2: In PySpark, create a program that reads a CSV file containing sales data, performs data cleaning by handling missing values and removing duplicates, calculates the total sales amount for each product, and finally, outputs the results to a new CSV file. Ensure to use transformations and actions in your PySpark script.

Steps we will be doing are :

1. Read the CSV file containing sales data.
2. Perform data cleaning: - Handle missing values. - Remove duplicate entries.
3. Calculate the total sales amount for each product.
4. Output the results to a new CSV file.



```
[5]: import pyspark

[6]: from pyspark.sql import SparkSession

[7]: spark=SparkSession.builder.appName('Practise').getOrCreate()

24/03/17 15:15:32 WARN Utils: Your hostname, Roshans-MacBook-Air.local resolves to a loopback address: 127.0.0.1; using 192.168.0.157 instead (on interface en0)
24/03/17 15:15:32 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).
24/03/17 15:15:32 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

[8]: spark

[8]: SparkSession - in-memory

     SparkContext

     Spark UI

     Version    v3.5.1
     Master     local[*]
     AppName    Practise

[55]: df_pyspark=spark.read.csv('sales_data_sample.csv',header=True,inferSchema=True)

[56]: df_pyspark

[56]: DataFrame[ORDERNUMBER: int, QUANTITYORDERED: int, PRICEEACH: double, ORDERLINENUMBER: int, SALES: double, ORDERDATE: string, STATUS: string, QTR_ID: int, MONTH_ID: int, YEAR_ID: int, PRODUCTLINE: string, MSRP: int, PRODUCTCODE: string, CUSTOMERNAME: string, PHONE: string, ADDRESSLINE1: string, ADDRESSLINE2: string, CITY: string, STATE: string, POSTALCODE: string, COUNTRY: string, TERRITORY: string, CONTACTLASTNAME: string, CONTACTFIRSTNAME: string, DEALSIZE: string]

[57]: df_pyspark.show()
```

[57]:
```
df_pyspark.show()
```

```
+-----------+---------------+---------+---------------+-------+---------------+-------+------+--------+-------+-----------+----+----------
-+--------------------+----------------+--------------------+------------+-------------+--------+----------+---------+---------+----------
-----+----------------+--------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES|      ORDERDATE| STATUS|QTR_ID|MONTH_ID|YEAR_ID|PRODUCTLINE|MSRP|PRODUCTCOD
E|        CUSTOMERNAME|           PHONE|        ADDRESSLINE1|ADDRESSLINE2|         CITY|   STATE|POSTALCODE|  COUNTRY|TERRITORY|CONTACTLAS
TNAME|CONTACTFIRSTNAME|DEALSIZE|
+-----------+---------------+---------+---------------+-------+---------------+-------+------+--------+-------+-----------+----+----------
-+--------------------+----------------+--------------------+------------+-------------+--------+----------+---------+---------+----------
-----+----------------+--------+
|      10107|             30|     95.7|              2| 2871.0| 2/24/2003 0:00|Shipped|     1|       2|   2003|Motorcycles|  95|   S10_167
8|   Land of Toys Inc.|      2125557818|897 Long Airport ...|        NULL|          NYC|      NY|     10022|      USA|       NA|
Yu|            Kwai|   Small|
|      10121|             34|    81.35|              5| 2765.9|  5/7/2003 0:00|Shipped|     2|       5|   2003|Motorcycles|  95|   S10_167
8|  Reims Collectables|      26.47.1555|  59 rue de l'Abbaye|        NULL|        Reims|    NULL|     51100|   France|     EMEA|        He
nriot|            Paul|   Small|
|      10134|             41|    94.74|              2|3884.34|  7/1/2003 0:00|Shipped|     3|       7|   2003|Motorcycles|  95|   S10_167
8|     Lyon Souveniers|+33 1 46 62 7555|27 rue du Colonel...|        NULL|        Paris|    NULL|     75508|   France|     EMEA|       Da
Cunha|          Daniel|  Medium|
|      10145|             45|    83.26|              6| 3746.7| 8/25/2003 0:00|Shipped|     3|       8|   2003|Motorcycles|  95|   S10_167
8|   Toys4GrownUps.com|      6265557265|  78934 Hillside Dr.|        NULL|     Pasadena|      CA|     90003|      USA|       NA|
Young|           Julie|  Medium|
|      10159|             49|    100.0|             14|5205.27|10/10/2003 0:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_167
8|Corporate Gift Id...|      6505551386|     7734 Strong St.|        NULL|San Francisco|      CA|      NULL|      USA|       NA|
Brown|           Julie|  Medium|
|      10168|             36|    96.66|              1|3479.76|10/28/2003 0:00|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_167
8|Technics Stores Inc.|      6505556809|   9408 Furth Circle|        NULL|   Burlingame|      CA|     94217|      USA|       NA|         H
irano|            Juri|  Medium|
|      10180|             29|    86.13|              9|2497.77|11/11/2003 0:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_167
8|Daedalus Designs ...|      20.16.1555|184, chausse de T...|        NULL|        Lille|    NULL|     59000|   France|     EMEA|
Rance|         Martine|   Small|
|      10188|             48|    100.0|              1|5512.32|11/18/2003 0:00|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_167
8|        Herkku Gifts|   +47 2267 3215|Drammen 121, PR 7...|        NULL|       Bergen|    NULL|    N 5804|   Norway|     EMEA|         O
eztan|          Veysel|  Medium|
|      10201|             22|    98.57|              2|2168.54| 12/1/2003 0:00|Shipped|     4|      12|   2003|Motorcycles|  95|   S10_167
8|     Mini Wheels Co.|      6505555787|5557 North Pendal...|        NULL|San Francisco|      CA|      NULL|      USA|       NA|         M
urphy|           Julie|   Small|
|      10211|             41|    100.0|             14|4708.44| 1/15/2004 0:00|Shipped|     1|       1|   2004|Motorcycles|  95|   S10_167
8|    Auto Canal Petit|  (1) 47.55.6555|   25, rue Lauriston|        NULL|        Paris|    NULL|     75016|   France|     EMEA|        Pe
rrier|       Dominique|  Medium|
|      10223|             37|    100.0|              1|3965.66| 2/20/2004 0:00|Shipped|     1|       2|   2004|Motorcycles|  95|   S10_167
8|Australian Collec...|    03 9520 4555|   636 St Kilda Road|     Level 3|    Melbourne|Victoria|      3004|Australia|     APAC|       Fer
guson|           Peter|  Medium|
|      10237|             23|    100.0|              7|2333.12|  4/5/2004 0:00|Shipped|     2|       4|   2004|Motorcycles|  95|   S10_167
8|     Vitachrome Inc.|      2125551500|   2678 Kingston Rd.|   Suite 101|          NYC|      NY|     10022|      USA|       NA|
Frick|         Michael|   Small|
|      10251|             28|    100.0|              2|3188.64| 5/18/2004 0:00|Shipped|     2|       5|   2004|Motorcycles|  95|   S10_167
8|Tekni Collectable...|      2015559350|       7476 Moss Rd.|        NULL|       Newark|      NJ|     94019|      USA|       NA|
Brown|         William|  Medium|
|      10263|             34|    100.0|              2|3676.76| 6/28/2004 0:00|Shipped|     2|       6|   2004|Motorcycles|  95|   S10_167
8|     Gift Depot Inc.|      2035552570| 25593 South Bay Ln.|        NULL|  Bridgewater|      CT|     97562|      USA|       NA|
King|           Julie|  Medium|
|      10275|             45|    92.83|              1|4177.35| 7/23/2004 0:00|Shipped|     3|       7|   2004|Motorcycles|  95|   S10_167
8|   La Rochelle Gifts|      40.67.8555|67, rue des Cinqu...|        NULL|       Nantes|    NULL|     44000|   France|     EMEA|        La
brune|          Janine|  Medium|
|      10285|             36|    100.0|              6|4099.68| 8/27/2004 0:00|Shipped|     3|       8|   2004|Motorcycles|  95|   S10_167
8|Marta's Replicas Co.|      6175558555| 39323 Spinnaker Dr.|        NULL|    Cambridge|      MA|     51247|      USA|       NA|      Hern
andez|           Marta|  Medium|
|      10299|             23|    100.0|              9|2597.39| 9/30/2004 0:00|Shipped|     3|       9|   2004|Motorcycles|  95|   S10_167
8|Toys of Finland, Co.|     90-224 8555|       Keskuskatu 45|        NULL|     Helsinki|    NULL|     21240|  Finland|     EMEA|      Kart
tunen|           Matti|   Small|
|      10309|             41|    100.0|              5|4394.38|10/15/2004 0:00|Shipped|     4|      10|   2004|Motorcycles|  95|   S10_167
8|  Baane Mini Imports|      07-98 9555|Erling Skakkes ga...|        NULL|      Stavern|    NULL|      4110|   Norway|     EMEA|     Bergu
lfsen|           Jonas|  Medium|
|      10318|             46|    94.74|              1|4358.04| 11/2/2004 0:00|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_167
8|Diecast Classics ...|      2155551555|    7586 Pompton St.|        NULL|    Allentown|      PA|     70267|      USA|       NA|
Yu|           Kyung|  Medium|
|      10329|             42|    100.0|              1|4396.14|11/15/2004 0:00|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_167
8|   Land of Toys Inc.|      2125557818|897 Long Airport ...|        NULL|          NYC|      NY|     10022|      USA|       NA|
Yu|            Kwai|  Medium|
+-----------+---------------+---------+---------------+-------+---------------+-------+------+--------+-------+-----------+----+----------
-+--------------------+----------------+--------------------+------------+-------------+--------+----------+---------+---------+----------
-----+----------------+--------+
only showing top 20 rows
```

[58]:
```
df_pyspark.head(3)
```

[58]:
```
[Row(ORDERNUMBER=10107, QUANTITYORDERED=30, PRICEEACH=95.7, ORDERLINENUMBER=2, SALES=2871.0, ORDERDATE='2/24/2003 0:00', STATUS='Shipped',
QTR_ID=1, MONTH_ID=2, YEAR_ID=2003, PRODUCTLINE='Motorcycles', MSRP=95, PRODUCTCODE='S10_1678', CUSTOMERNAME='Land of Toys Inc.', PHONE='2
125557818', ADDRESSLINE1='897 Long Airport Avenue', ADDRESSLINE2=None, CITY='NYC', STATE='NY', POSTALCODE='10022', COUNTRY='USA', TERRITOR
Y='NA', CONTACTLASTNAME='Yu', CONTACTFIRSTNAME='Kwai', DEALSIZE='Small'),
 Row(ORDERNUMBER=10121, QUANTITYORDERED=34, PRICEEACH=81.35, ORDERLINENUMBER=5, SALES=2765.9, ORDERDATE='5/7/2003 0:00', STATUS='Shipped',
QTR_ID=2, MONTH_ID=5, YEAR_ID=2003, PRODUCTLINE='Motorcycles', MSRP=95, PRODUCTCODE='S10_1678', CUSTOMERNAME='Reims Collectables', PHONE
='26.47.1555', ADDRESSLINE1="59 rue de l'Abbaye", ADDRESSLINE2=None, CITY='Reims', STATE=None, POSTALCODE='51100', COUNTRY='France', TERRI
TORY='EMEA', CONTACTLASTNAME='Henriot', CONTACTFIRSTNAME='Paul', DEALSIZE='Small'),
 Row(ORDERNUMBER=10134, QUANTITYORDERED=41, PRICEEACH=94.74, ORDERLINENUMBER=2, SALES=3884.34, ORDERDATE='7/1/2003 0:00', STATUS='Shippe
d', QTR_ID=3, MONTH_ID=7, YEAR_ID=2003, PRODUCTLINE='Motorcycles', MSRP=95, PRODUCTCODE='S10_1678', CUSTOMERNAME='Lyon Souveniers', PHONE
='+33 1 46 62 7555', ADDRESSLINE1='27 rue du Colonel Pierre Avia', ADDRESSLINE2=None, CITY='Paris', STATE=None, POSTALCODE='75508', COUNTR
Y='France', TERRITORY='EMEA', CONTACTLASTNAME='Da Cunha', CONTACTFIRSTNAME='Daniel', DEALSIZE='Medium')]
```

[59]:
```
df_pyspark=df_pyspark.drop('ORDERDATE')
```

[60]:
```
df_pyspark
```

[60]:
```
DataFrame[ORDERNUMBER: int, QUANTITYORDERED: int, PRICEEACH: double, ORDERLINENUMBER: int, SALES: double, STATUS: string, QTR_ID: int, MON
TH_ID: int, YEAR_ID: int, PRODUCTLINE: string, MSRP: int, PRODUCTCODE: string, CUSTOMERNAME: string, PHONE: string, ADDRESSLINE1: string,
ADDRESSLINE2: string, CITY: string, STATE: string, POSTALCODE: string, COUNTRY: string, TERRITORY: string, CONTACTLASTNAME: string, CONTAC
TFIRSTNAME: string, DEALSIZE: string]
```

[61]:
```
df_pyspark.show()
```

TFIRSTNAME: string, DEALSIZE: string]

[61]: df_pyspark.show()

```
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+-----------+----+-----------+--------------
------+---------------+--------------------+------------+-------------+--------+----------+---------+---------+--------------+----------
------+--------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES| STATUS|QTR_ID|MONTH_ID|YEAR_ID|PRODUCTLINE|MSRP|PRODUCTCODE|        CUSTOM
ERNAME|          PHONE|        ADDRESSLINE1|ADDRESSLINE2|         CITY|   STATE|POSTALCODE|  COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIR
STNAME|DEALSIZE|
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+-----------+----+-----------+--------------
------+---------------+--------------------+------------+-------------+--------+----------+---------+---------+--------------+----------
------+--------+
|      10107|             30|     95.7|              2| 2871.0|Shipped|     1|       2|   2003|Motorcycles|  95|   S10_1678|   Land of Toy
s Inc.|     2125557818|897 Long Airport ...|        NULL|          NYC|      NY|     10022|      USA|       NA|             Yu|
Kwai|   Small|
|      10121|             34|    81.35|              5| 2765.9|Shipped|     2|       5|   2003|Motorcycles|  95|   S10_1678|  Reims Collec
tables|     26.47.1555|  59 rue de l'Abbaye|        NULL|        Reims|    NULL|     51100|   France|     EMEA|        Henriot|
Paul|   Small|
|      10134|             41|    94.74|              2|3884.34|Shipped|     3|       7|   2003|Motorcycles|  95|   S10_1678|     Lyon Souv
eniers|+33 1 46 62 7555|27 rue du Colonel...|        NULL|        Paris|    NULL|     75508|   France|     EMEA|       Da Cunha|
Daniel|  Medium|
|      10145|             45|    83.26|              6| 3746.7|Shipped|     3|       8|   2003|Motorcycles|  95|   S10_1678|   Toys4GrownU
ps.com|     6265557265|  78934 Hillside Dr.|        NULL|     Pasadena|      CA|     90003|      USA|       NA|          Young|
Julie|  Medium|
|      10159|             49|    100.0|             14|5205.27|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Corporate Gift
Id...|     6505551386|     7734 Strong St.|        NULL|San Francisco|      CA|      NULL|      USA|       NA|          Brown|
Julie|  Medium|
|      10168|             36|    96.66|              1|3479.76|Shipped|     4|      10|   2003|Motorcycles|  95|   S10_1678|Technics Store
s Inc.|     6505556809|   9408 Furth Circle|        NULL|   Burlingame|      CA|     94217|      USA|       NA|         Hirano|
Juri|  Medium|
|      10180|             29|    86.13|              9|2497.77|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|Daedalus Desig
ns ...|     20.16.1555|184, chausse de T...|        NULL|        Lille|    NULL|     59000|   France|     EMEA|          Rance|         M
artine|   Small|
|      10188|             48|    100.0|              1|5512.32|Shipped|     4|      11|   2003|Motorcycles|  95|   S10_1678|        Herkku
Gifts|   +47 2267 3215|Drammen 121, PR 7...|        NULL|       Bergen|    NULL|    N 5804|   Norway|     EMEA|         Oeztan|         V
eysel|  Medium|
|      10201|             22|    98.57|              2|2168.54|Shipped|     4|      12|   2003|Motorcycles|  95|   S10_1678|     Mini Whee
ls Co.|     6505555787|5557 North Pendal...|        NULL|San Francisco|      CA|      NULL|      USA|       NA|         Murphy|
Julie|   Small|
|      10211|             41|    100.0|             14|4708.44|Shipped|     1|       1|   2004|Motorcycles|  95|   S10_1678|    Auto Canal
Petit|  (1) 47.55.6555|   25, rue Lauriston|        NULL|        Paris|    NULL|     75016|   France|     EMEA|        Perrier|      Domi
nique|  Medium|
|      10223|             37|    100.0|              1|3965.66|Shipped|     1|       2|   2004|Motorcycles|  95|   S10_1678|Australian Col
lec...|    03 9520 4555|   636 St Kilda Road|     Level 3|    Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10237|             23|    100.0|              7|2333.12|Shipped|     2|       4|   2004|Motorcycles|  95|   S10_1678|     Vitachrom
e Inc.|     2125551500|   2678 Kingston Rd.|   Suite 101|          NYC|      NY|     10022|      USA|       NA|          Frick|         M
ichael|   Small|
|      10251|             28|    100.0|              2|3188.64|Shipped|     2|       5|   2004|Motorcycles|  95|   S10_1678|Tekni Collecta
ble...|     2015559350|       7476 Moss Rd.|        NULL|       Newark|      NJ|     94019|      USA|       NA|          Brown|         W
illiam|  Medium|
|      10263|             34|    100.0|              2|3676.76|Shipped|     2|       6|   2004|Motorcycles|  95|   S10_1678|     Gift Depo
t Inc.|     2035552570| 25593 South Bay Ln.|        NULL|  Bridgewater|      CT|     97562|      USA|       NA|           King|
Julie|  Medium|
|      10275|             45|    92.83|              1|4177.35|Shipped|     3|       7|   2004|Motorcycles|  95|   S10_1678|   La Rochelle
Gifts|     40.67.8555|67, rue des Cinqu...|        NULL|       Nantes|    NULL|     44000|   France|     EMEA|        Labrune|         J
anine|  Medium|
|      10285|             36|    100.0|              6|4099.68|Shipped|     3|       8|   2004|Motorcycles|  95|   S10_1678|Marta's Replic
as Co.|     6175558555| 39323 Spinnaker Dr.|        NULL|    Cambridge|      MA|     51247|      USA|       NA|      Hernandez|
Marta|  Medium|
|      10299|             23|    100.0|              9|2597.39|Shipped|     3|       9|   2004|Motorcycles|  95|   S10_1678|Toys of Finlan
d, Co.|    90-224 8555|       Keskuskatu 45|        NULL|     Helsinki|    NULL|     21240|  Finland|     EMEA|      Karttunen|
Matti|   Small|
|      10309|             41|    100.0|              5|4394.38|Shipped|     4|      10|   2004|Motorcycles|  95|   S10_1678|  Baane Mini I
mports|     07-98 9555|Erling Skakkes ga...|        NULL|      Stavern|    NULL|      4110|   Norway|     EMEA|      Bergulfsen|
Jonas|  Medium|
|      10318|             46|    94.74|              1|4358.04|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_1678|Diecast Classi
cs ...|     2155551555|    7586 Pompton St.|        NULL|    Allentown|      PA|     70267|      USA|       NA|             Yu|
Kyung|  Medium|
|      10329|             42|    100.0|              1|4396.14|Shipped|     4|      11|   2004|Motorcycles|  95|   S10_1678|   Land of Toy
s Inc.|     2125557818|897 Long Airport ...|        NULL|          NYC|      NY|     10022|      USA|       NA|             Yu|
Kwai|  Medium|
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+-----------+----+-----------+--------------
------+---------------+--------------------+------------+-------------+--------+----------+---------+---------+--------------+----------
------+--------+
only showing top 20 rows
```

[62]:
```python
df_pyspark = df_pyspark.na.drop(how="any")
```

[63]:
```python
df_pyspark.show()
```

```
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+------------+----+-----------+-------------
-------+---------------+--------------------+------------+-------------+---------+----------+---------+---------+---------------+-----------
-----+--------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES| STATUS|QTR_ID|MONTH_ID|YEAR_ID| PRODUCTLINE|MSRP|PRODUCTCODE|        CUSTO
MERNAME|          PHONE|        ADDRESSLINE1|ADDRESSLINE2|        CITY|    STATE|POSTALCODE|  COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRS
TNAME|DEALSIZE|
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+------------+----+-----------+-------------
-------+---------------+--------------------+------------+-------------+---------+----------+---------+---------+---------------+-----------
-----+--------+
|      10223|             37|    100.0|              1|3965.66|Shipped|     1|       2|   2004| Motorcycles|  95|   S10_1678|Australian Co
llec...|   03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10237|             23|    100.0|              7|2333.12|Shipped|     2|       4|   2004| Motorcycles|  95|   S10_1678|     Vitachro
me Inc.|      2125551500|   2678 Kingston Rd.|   Suite 101|         NYC|       NY|     10022|      USA|       NA|          Frick|          Mi
chael|   Small|
|      10361|             20|    72.55|             13| 1451.0|Shipped|     4|      12|   2004| Motorcycles|  95|   S10_1678|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|      NSW|      2067|Australia|     APAC|         Huxley|           A
drian|   Small|
|      10163|             21|    100.0|              1|4860.24|Shipped|     4|      10|   2003|Classic Cars| 214|   S10_1949|Classic Legen
ds Inc.|      2125558493|    5905 Pompton St.|   Suite 750|         NYC|       NY|     10022|      USA|       NA|      Hernandez|
Maria|  Medium|
|      10270|             21|    100.0|              9|4905.39|Shipped|     3|       7|   2004|Classic Cars| 214|   S10_1949|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|      NSW|      2067|Australia|     APAC|         Huxley|           A
drian|  Medium|
|      10347|             30|    100.0|              1| 3944.7|Shipped|     4|      11|   2004|Classic Cars| 214|   S10_1949|Australian Co
llec...|   03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10391|             24|    100.0|              4|2416.56|Shipped|     1|       3|   2005|Classic Cars| 214|   S10_1949|Anna's Decora
tion...|   02 9936 8555|   201 Miller Street|    Level 15|North Sydney|      NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|
|      10120|             29|    96.34|              3|2793.86|Shipped|     2|       4|   2003| Motorcycles| 118|   S10_2016|Australian Co
llec...|   03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Small|
|      10223|             47|    100.0|              4|5422.39|Shipped|     1|       2|   2004| Motorcycles| 118|   S10_2016|Australian Co
llec...|   03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10361|             26|    51.15|              8| 1329.9|Shipped|     4|      12|   2004| Motorcycles| 118|   S10_2016|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|      NSW|      2067|Australia|     APAC|         Huxley|           A
drian|   Small|
|      10120|             46|    100.0|              2|9264.86|Shipped|     2|       4|   2003| Motorcycles| 193|   S10_4698|Australian Co
llec...|   03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|
|      10223|             49|    100.0|              3|9774.03|Shipped|     1|       2|   2004| Motorcycles| 193|   S10_4698|Australian Co
llec...|   03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|
|      10237|             39|    100.0|              9| 7023.9|Shipped|     2|       4|   2004| Motorcycles| 193|   S10_4698|     Vitachro
me Inc.|      2125551500|   2678 Kingston Rd.|   Suite 101|         NYC|       NY|     10022|      USA|       NA|          Frick|          Mi
chael|   Large|
|      10337|             25|    48.05|              8|1201.25|Shipped|     4|      11|   2004|Classic Cars| 136|   S10_4757|Classic Legen
ds Inc.|      2125558493|    5905 Pompton St.|   Suite 750|         NYC|       NY|     10022|      USA|       NA|      Hernandez|
Maria|   Small|
|      10270|             32|    100.0|              2|4302.08|Shipped|     3|       7|   2004|Classic Cars| 147|   S10_4962|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|      NSW|      2067|Australia|     APAC|         Huxley|           A
drian|  Medium|
|      10347|             27|    100.0|              2| 4428.0|Shipped|     4|      11|   2004|Classic Cars| 147|   S10_4962|Australian Co
llec...|   03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10370|             35|    65.63|              4|2297.05|Shipped|     1|       1|   2005|Classic Cars| 147|   S10_4962|Anna's Decora
tion...|   02 9936 8555|   201 Miller Street|    Level 15|North Sydney|      NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|
|      10391|             37|     46.9|              7| 1735.3|Shipped|     1|       3|   2005|Classic Cars| 147|   S10_4962|Anna's Decora
tion...|   02 9936 8555|   201 Miller Street|    Level 15|North Sydney|      NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|
|      10169|             30|    100.0|              2| 5019.9|Shipped|     4|      11|   2003|Classic Cars| 194|   S12_1099|Anna's Decora
tion...|   02 9936 8555|   201 Miller Street|    Level 15|North Sydney|      NSW|      2060|Australia|     APAC|         O'Hara|
Anna|  Medium|
|      10127|             46|    100.0|              2|11279.2|Shipped|     2|       6|   2003|Classic Cars| 207|   S12_1108|  Muscle Mach
ine Inc|      2125557413|    4092 Furth Circle|   Suite 400|         NYC|       NY|     10022|      USA|       NA|          Young|
Jeff|   Large|
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+------------+----+-----------+-------------
-------+---------------+--------------------+------------+-------------+---------+----------+---------+---------+---------------+-----------
-----+--------+
```

[64]:
```
df_pyspark=df_pyspark.na.drop(how="any",subset=['COUNTRY'])
```

[65]:
```
df_pyspark.show()
```

```
+-----------+---------------+---------+---------------+-------+-------+-------+--------+-------+------------+----+-----------+------------
-------+---------------+--------------------+------------+------------+---------+----------+---------+---------+---------------+-----------
-----+--------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES| STATUS|QTR_ID|MONTH_ID|YEAR_ID| PRODUCTLINE|MSRP|PRODUCTCODE|        CUSTO
MERNAME|          PHONE|        ADDRESSLINE1|ADDRESSLINE2|        CITY|   STATE|POSTALCODE|  COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRS
TNAME|DEALSIZE|
+-----------+---------------+---------+---------------+-------+-------+-------+--------+-------+------------+----+-----------+------------
-------+---------------+--------------------+------------+------------+---------+----------+---------+---------+---------------+-----------
-----+--------+
|      10223|             37|    100.0|              1|3965.66|Shipped|     1|       2|   2004| Motorcycles|  95|   S10_1678|Australian Co
llec...|  03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10237|             23|    100.0|              7|2333.12|Shipped|     2|       4|   2004| Motorcycles|  95|   S10_1678|     Vitachro
me Inc.|     2125551500|   2678 Kingston Rd.|   Suite 101|         NYC|      NY|     10022|      USA|       NA|          Frick|          Mi
chael|   Small|
|      10361|             20|    72.55|             13| 1451.0|Shipped|     4|      12|   2004| Motorcycles|  95|   S10_1678|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|           A
drian|   Small|
|      10163|             21|    100.0|              1|4860.24|Shipped|     4|      10|   2003|Classic Cars| 214|   S10_1949|Classic Legen
ds Inc.|     2125558493|    5905 Pompton St.|   Suite 750|         NYC|      NY|     10022|      USA|       NA|      Hernandez|
Maria|  Medium|
|      10270|             21|    100.0|              9|4905.39|Shipped|     3|       7|   2004|Classic Cars| 214|   S10_1949|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|           A
drian|  Medium|
|      10347|             30|    100.0|              1| 3944.7|Shipped|     4|      11|   2004|Classic Cars| 214|   S10_1949|Australian Co
llec...|  03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10391|             24|    100.0|              4|2416.56|Shipped|     1|       3|   2005|Classic Cars| 214|   S10_1949|Anna's Decora
tion...|  02 9936 8555|   201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|
|      10120|             29|    96.34|              3|2793.86|Shipped|     2|       4|   2003| Motorcycles| 118|   S10_2016|Australian Co
llec...|  03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Small|
|      10223|             47|    100.0|              4|5422.39|Shipped|     1|       2|   2004| Motorcycles| 118|   S10_2016|Australian Co
llec...|  03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10361|             26|    51.15|              8| 1329.9|Shipped|     4|      12|   2004| Motorcycles| 118|   S10_2016|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|           A
drian|   Small|
|      10120|             46|    100.0|              2|9264.86|Shipped|     2|       4|   2003| Motorcycles| 193|   S10_4698|Australian Co
llec...|  03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|
|      10223|             49|    100.0|              3|9774.03|Shipped|     1|       2|   2004| Motorcycles| 193|   S10_4698|Australian Co
llec...|  03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|
|      10237|             39|    100.0|              9| 7023.9|Shipped|     2|       4|   2004| Motorcycles| 193|   S10_4698|     Vitachro
me Inc.|     2125551500|   2678 Kingston Rd.|   Suite 101|         NYC|      NY|     10022|      USA|       NA|          Frick|          Mi
chael|   Large|
|      10337|             25|    48.05|              8|1201.25|Shipped|     4|      11|   2004|Classic Cars| 136|   S10_4757|Classic Legen
ds Inc.|     2125558493|    5905 Pompton St.|   Suite 750|         NYC|      NY|     10022|      USA|       NA|      Hernandez|
Maria|   Small|
|      10270|             32|    100.0|              2|4302.08|Shipped|     3|       7|   2004|Classic Cars| 147|   S10_4962|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|           A
drian|  Medium|
|      10347|             27|    100.0|              2| 4428.0|Shipped|     4|      11|   2004|Classic Cars| 147|   S10_4962|Australian Co
llec...|  03 9520 4555|   636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10370|             35|    65.63|              4|2297.05|Shipped|     1|       1|   2005|Classic Cars| 147|   S10_4962|Anna's Decora
tion...|  02 9936 8555|   201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|
|      10391|             37|     46.9|              7| 1735.3|Shipped|     1|       3|   2005|Classic Cars| 147|   S10_4962|Anna's Decora
tion...|  02 9936 8555|   201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|
|      10169|             30|    100.0|              2| 5019.9|Shipped|     4|      11|   2003|Classic Cars| 194|   S12_1099|Anna's Decora
tion...|  02 9936 8555|   201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|  Medium|
|      10127|             46|    100.0|              2|11279.2|Shipped|     2|       6|   2003|Classic Cars| 207|   S12_1108|  Muscle Mach
ine Inc|     2125557413|   4092 Furth Circle|   Suite 400|         NYC|      NY|     10022|      USA|       NA|          Young|
Jeff|   Large|
+-----------+---------------+---------+---------------+-------+-------+-------+--------+-------+------------+----+-----------+------------
-------+---------------+--------------------+------------+------------+---------+----------+---------+---------+---------------+-----------
-----+--------+
```

[66]:

```python
from pyspark.ml.feature import Imputer

imputer = Imputer(
    inputCols=['QUANTITYORDERED','PRICEEACH', 'SALES'],
    outputCols=["{}_imputed".format(c) for c in ['QUANTITYORDERED','PRICEEACH', 'SALES']]
    ).setStrategy("median")
```

[67]:

```python
imputer.fit(df_pyspark).transform(df_pyspark).show()
```

```
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+------------+----+-----------+------------
-------+--------------+--------------------+------------+------------+--------+----------+---------+---------+---------------+----------
-----+--------+-----------------------+-----------------+-------------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES| STATUS|QTR_ID|MONTH_ID|YEAR_ID| PRODUCTLINE|MSRP|PRODUCTCODE|        CUSTO
MERNAME|         PHONE|        ADDRESSLINE1|ADDRESSLINE2|        CITY|   STATE|POSTALCODE|  COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRS
TNAME|DEALSIZE|QUANTITYORDERED_imputed|PRICEEACH_imputed|SALES_imputed|
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+------------+----+-----------+------------
-------+--------------+--------------------+------------+------------+--------+----------+---------+---------+---------------+----------
-----+--------+-----------------------+-----------------+-------------+
|      10223|             37|    100.0|              1|3965.66|Shipped|     1|       2|   2004| Motorcycles|  95|   S10_1678|Australian Co
llec...|  03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|                     37|            100.0|      3965.66|
|      10237|             23|    100.0|              7|2333.12|Shipped|     2|       4|   2004| Motorcycles|  95|   S10_1678|     Vitachro
me Inc.|      2125551500|  2678 Kingston Rd.|   Suite 101|         NYC|      NY|     10022|      USA|       NA|          Frick|         Mi
chael|   Small|                     23|            100.0|      2333.12|
|      10361|             20|    72.55|             13| 1451.0|Shipped|     4|      12|   2004| Motorcycles|  95|   S10_1678|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|          A
drian|   Small|                     20|            72.55|       1451.0|
|      10163|             21|    100.0|              1|4860.24|Shipped|     4|      10|   2003|Classic Cars| 214|   S10_1949|Classic Legen
ds Inc.|      2125558493|    5905 Pompton St.|   Suite 750|         NYC|      NY|     10022|      USA|       NA|      Hernandez|
Maria|  Medium|                     21|            100.0|      4860.24|
|      10270|             21|    100.0|              9|4905.39|Shipped|     3|       7|   2004|Classic Cars| 214|   S10_1949|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|          A
drian|  Medium|                     21|            100.0|      4905.39|
|      10347|             30|    100.0|              1| 3944.7|Shipped|     4|      11|   2004|Classic Cars| 214|   S10_1949|Australian Co
llec...|  03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|                     30|            100.0|       3944.7|
|      10391|             24|    100.0|              4|2416.56|Shipped|     1|       3|   2005|Classic Cars| 214|   S10_1949|Anna's Decora
tion...|  02 9936 8555|  201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|                     24|            100.0|      2416.56|
|      10120|             29|    96.34|              3|2793.86|Shipped|     2|       4|   2003| Motorcycles| 118|   S10_2016|Australian Co
llec...|  03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Small|                     29|            96.34|      2793.86|
|      10223|             47|    100.0|              4|5422.39|Shipped|     1|       2|   2004| Motorcycles| 118|   S10_2016|Australian Co
llec...|  03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|                     47|            100.0|      5422.39|
|      10361|             26|    51.15|              8| 1329.9|Shipped|     4|      12|   2004| Motorcycles| 118|   S10_2016|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|          A
drian|   Small|                     26|            51.15|       1329.9|
|      10120|             46|    100.0|              2|9264.86|Shipped|     2|       4|   2003| Motorcycles| 193|   S10_4698|Australian Co
llec...|  03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|                     46|            100.0|      9264.86|
|      10223|             49|    100.0|              3|9774.03|Shipped|     1|       2|   2004| Motorcycles| 193|   S10_4698|Australian Co
llec...|  03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|                     49|            100.0|      9774.03|
|      10237|             39|    100.0|              9| 7023.9|Shipped|     2|       4|   2004| Motorcycles| 193|   S10_4698|     Vitachro
me Inc.|      2125551500|  2678 Kingston Rd.|   Suite 101|         NYC|      NY|     10022|      USA|       NA|          Frick|         Mi
chael|   Large|                     39|            100.0|       7023.9|
|      10337|             25|    48.05|              8|1201.25|Shipped|     4|      11|   2004|Classic Cars| 136|   S10_4757|Classic Legen
ds Inc.|      2125558493|    5905 Pompton St.|   Suite 750|         NYC|      NY|     10022|      USA|       NA|      Hernandez|
Maria|   Small|                     25|            48.05|      1201.25|
|      10270|             32|    100.0|              2|4302.08|Shipped|     3|       7|   2004|Classic Cars| 147|   S10_4962|Souveniers An
d Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|          A
drian|  Medium|                     32|            100.0|      4302.08|
|      10347|             27|    100.0|              2| 4428.0|Shipped|     4|      11|   2004|Classic Cars| 147|   S10_4962|Australian Co
llec...|  03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|                     27|            100.0|       4428.0|
|      10370|             35|    65.63|              4|2297.05|Shipped|     1|       1|   2005|Classic Cars| 147|   S10_4962|Anna's Decora
tion...|  02 9936 8555|  201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|                     35|            65.63|      2297.05|
|      10391|             37|     46.9|              7| 1735.3|Shipped|     1|       3|   2005|Classic Cars| 147|   S10_4962|Anna's Decora
tion...|  02 9936 8555|  201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|                     37|             46.9|       1735.3|
|      10169|             30|    100.0|              2| 5019.9|Shipped|     4|      11|   2003|Classic Cars| 194|   S12_1099|Anna's Decora
tion...|  02 9936 8555|  201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|  Medium|                     30|            100.0|       5019.9|
|      10127|             46|    100.0|              2|11279.2|Shipped|     2|       6|   2003|Classic Cars| 207|   S12_1108|  Muscle Mach
```

justed by setting `spark.sql.debug.maxToStringFields`.

[68]:
```python
df_pyspark.show()
```

```
+-----------+---------------+---------+---------------+-------+-------+-------+--------+-------+------------+----+-----------+---------
-----------+---------------+--------------------+------------+-------------+--------+----------+---------+-----------+---------------+---
-------------+--------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES| STATUS|QTR_ID|MONTH_ID|YEAR_ID| PRODUCTLINE|MSRP|PRODUCTCODE|        C
USTOMERNAME|          PHONE|        ADDRESSLINE1|ADDRESSLINE2|         CITY|   STATE|POSTALCODE|  COUNTRY|TERRITORY|CONTACTLASTNAME|CON
TACTFIRSTNAME|DEALSIZE|
+-----------+---------------+---------+---------------+-------+-------+-------+--------+-------+------------+----+-----------+---------
-----------+---------------+--------------------+------------+-------------+--------+----------+---------+-----------+---------------+---
-------------+--------+
|      10223|             37|    100.0|              1|3965.66|Shipped|     1|       2|   2004| Motorcycles|  95|   S10_1678|Australia
n Collec...|   03 9520 4555|   636 St Kilda Road|     Level 3|    Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10237|             23|    100.0|              7|2333.12|Shipped|     2|       4|   2004| Motorcycles|  95|   S10_1678|     Vita
chrome Inc.|     2125551500|  2678 Kingston Rd.|   Suite 101|          NYC|      NY|     10022|      USA|       NA|          Frick|
Michael|   Small|
|      10361|             20|    72.55|             13| 1451.0|Shipped|     4|      12|   2004| Motorcycles|  95|   S10_1678|Souvenier
s And Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|    Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|
Adrian|   Small|
|      10163|             21|    100.0|              1|4860.24|Shipped|     4|      10|   2003|Classic Cars| 214|   S10_1949|Classic L
egends Inc.|     2125558493|    5905 Pompton St.|   Suite 750|          NYC|      NY|     10022|      USA|       NA|      Hernandez|
Maria|  Medium|
|      10270|             21|    100.0|              9|4905.39|Shipped|     3|       7|   2004|Classic Cars| 214|   S10_1949|Souvenier
s And Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|    Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|
Adrian|  Medium|
|      10347|             30|    100.0|              1| 3944.7|Shipped|     4|      11|   2004|Classic Cars| 214|   S10_1949|Australia
n Collec...|   03 9520 4555|   636 St Kilda Road|     Level 3|    Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10391|             24|    100.0|              4|2416.56|Shipped|     1|       3|   2005|Classic Cars| 214|   S10_1949|Anna's De
coration...|   02 9936 8555|   201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|
|      10120|             29|    96.34|              3|2793.86|Shipped|     2|       4|   2003| Motorcycles| 118|   S10_2016|Australia
n Collec...|   03 9520 4555|   636 St Kilda Road|     Level 3|    Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Small|
|      10223|             47|    100.0|              4|5422.39|Shipped|     1|       2|   2004| Motorcycles| 118|   S10_2016|Australia
n Collec...|   03 9520 4555|   636 St Kilda Road|     Level 3|    Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10361|             26|    51.15|              8| 1329.9|Shipped|     4|      12|   2004| Motorcycles| 118|   S10_2016|Souvenier
s And Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|    Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|
Adrian|   Small|
|      10120|             46|    100.0|              2|9264.86|Shipped|     2|       4|   2003| Motorcycles| 193|   S10_4698|Australia
n Collec...|   03 9520 4555|   636 St Kilda Road|     Level 3|    Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|
|      10223|             49|    100.0|              3|9774.03|Shipped|     1|       2|   2004| Motorcycles| 193|   S10_4698|Australia
n Collec...|   03 9520 4555|   636 St Kilda Road|     Level 3|    Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|
|      10237|             39|    100.0|              9| 7023.9|Shipped|     2|       4|   2004| Motorcycles| 193|   S10_4698|     Vita
chrome Inc.|     2125551500|  2678 Kingston Rd.|   Suite 101|          NYC|      NY|     10022|      USA|       NA|          Frick|
Michael|   Large|
|      10337|             25|    48.05|              8|1201.25|Shipped|     4|      11|   2004|Classic Cars| 136|   S10_4757|Classic L
egends Inc.|     2125558493|    5905 Pompton St.|   Suite 750|          NYC|      NY|     10022|      USA|       NA|      Hernandez|
Maria|   Small|
|      10270|             32|    100.0|              2|4302.08|Shipped|     3|       7|   2004|Classic Cars| 147|   S10_4962|Souvenier
s And Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|    Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|
Adrian|  Medium|
|      10347|             27|    100.0|              2| 4428.0|Shipped|     4|      11|   2004|Classic Cars| 147|   S10_4962|Australia
n Collec...|   03 9520 4555|   636 St Kilda Road|     Level 3|    Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|
|      10370|             35|    65.63|              4|2297.05|Shipped|     1|       1|   2005|Classic Cars| 147|   S10_4962|Anna's De
coration...|   02 9936 8555|   201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|
|      10391|             37|     46.9|              7| 1735.3|Shipped|     1|       3|   2005|Classic Cars| 147|   S10_4962|Anna's De
coration...|   02 9936 8555|   201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|
|      10169|             30|    100.0|              2| 5019.9|Shipped|     4|      11|   2003|Classic Cars| 194|   S12_1099|Anna's De
coration...|   02 9936 8555|   201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|  Medium|
|      10127|             46|    100.0|              2|11279.2|Shipped|     2|       6|   2003|Classic Cars| 207|   S12_1108|   Muscle
Machine Inc|     2125557413|   4092 Furth Circle|   Suite 400|          NYC|      NY|     10022|      USA|       NA|          Young|
Jeff|   Large|
+-----------+---------------+---------+---------------+-------+-------+-------+--------+-------+------------+----+-----------+---------
-----------+---------------+--------------------+------------+-------------+--------+----------+---------+-----------+---------------+---
-------------+--------+
only showing top 20 rows
```

[69]:

```python
imputed_df = imputer.fit(df_pyspark).transform(df_pyspark)
imputed_df.show()
```

```
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+------------+----+-----------+--------
-----------+---------------+--------------------+-------------+------------+--------+----------+---------+---------+---------------+---
-------------+--------+----------------------+-----------------+-------------+
|ORDERNUMBER|QUANTITYORDERED|PRICEEACH|ORDERLINENUMBER|  SALES| STATUS|QTR_ID|MONTH_ID|YEAR_ID| PRODUCTLINE|MSRP|PRODUCTCODE|        C
USTOMERNAME|          PHONE|        ADDRESSLINE1|ADDRESSLINE2|        CITY|   STATE|POSTALCODE|  COUNTRY|TERRITORY|CONTACTLASTNAME|CON
TACTFIRSTNAME|DEALSIZE|QUANTITYORDERED_imputed|PRICEEACH_imputed|SALES_imputed|
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+------------+----+-----------+--------
-----------+---------------+--------------------+-------------+------------+--------+----------+---------+---------+---------------+---
-------------+--------+----------------------+-----------------+-------------+
|      10223|             37|    100.0|              1|3965.66|Shipped|     1|       2|   2004| Motorcycles|  95|   S10_1678|Australia
n Collec...|   03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|                     37|            100.0|      3965.66|
|      10237|             23|    100.0|              7|2333.12|Shipped|     2|       4|   2004| Motorcycles|  95|   S10_1678|     Vita
chrome Inc.|     2125551500|  2678 Kingston Rd.|   Suite 101|         NYC|      NY|     10022|      USA|       NA|          Frick|
Michael|   Small|                     23|            100.0|      2333.12|
|      10361|             20|    72.55|             13| 1451.0|Shipped|     4|      12|   2004| Motorcycles|  95|   S10_1678|Souvenier
s And Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|
Adrian|   Small|                     20|            72.55|       1451.0|
|      10163|             21|    100.0|              1|4860.24|Shipped|     4|      10|   2003|Classic Cars| 214|   S10_1949|Classic L
egends Inc.|     2125558493|    5905 Pompton St.|   Suite 750|         NYC|      NY|     10022|      USA|       NA|      Hernandez|
Maria|  Medium|                     21|            100.0|      4860.24|
|      10270|             21|    100.0|              9|4905.39|Shipped|     3|       7|   2004|Classic Cars| 214|   S10_1949|Souvenier
s And Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|
Adrian|  Medium|                     21|            100.0|      4905.39|
|      10347|             30|    100.0|              1| 3944.7|Shipped|     4|      11|   2004|Classic Cars| 214|   S10_1949|Australia
n Collec...|   03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|                     30|            100.0|       3944.7|
|      10391|             24|    100.0|              4|2416.56|Shipped|     1|       3|   2005|Classic Cars| 214|   S10_1949|Anna's De
coration...|   02 9936 8555|  201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|                     24|            100.0|      2416.56|
|      10120|             29|    96.34|              3|2793.86|Shipped|     2|       4|   2003| Motorcycles| 118|   S10_2016|Australia
n Collec...|   03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Small|                     29|            96.34|      2793.86|
|      10223|             47|    100.0|              4|5422.39|Shipped|     1|       2|   2004| Motorcycles| 118|   S10_2016|Australia
n Collec...|   03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|                     47|            100.0|      5422.39|
|      10361|             26|    51.15|              8| 1329.9|Shipped|     4|      12|   2004| Motorcycles| 118|   S10_2016|Souvenier
s And Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|
Adrian|   Small|                     26|            51.15|       1329.9|
|      10120|             46|    100.0|              2|9264.86|Shipped|     2|       4|   2003| Motorcycles| 193|   S10_4698|Australia
n Collec...|   03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|                     46|            100.0|      9264.86|
|      10223|             49|    100.0|              3|9774.03|Shipped|     1|       2|   2004| Motorcycles| 193|   S10_4698|Australia
n Collec...|   03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|   Large|                     49|            100.0|      9774.03|
|      10237|             39|    100.0|              9| 7023.9|Shipped|     2|       4|   2004| Motorcycles| 193|   S10_4698|     Vita
chrome Inc.|     2125551500|  2678 Kingston Rd.|   Suite 101|         NYC|      NY|     10022|      USA|       NA|          Frick|
Michael|   Large|                     39|            100.0|       7023.9|
|      10337|             25|    48.05|              8|1201.25|Shipped|     4|      11|   2004|Classic Cars| 136|   S10_4757|Classic L
egends Inc.|     2125558493|    5905 Pompton St.|   Suite 750|         NYC|      NY|     10022|      USA|       NA|      Hernandez|
Maria|   Small|                     25|            48.05|      1201.25|
|      10270|             32|    100.0|              2|4302.08|Shipped|     3|       7|   2004|Classic Cars| 147|   S10_4962|Souvenier
s And Th...|+61 2 9495 8555|Monitor Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|
Adrian|  Medium|                     32|            100.0|      4302.08|
|      10347|             27|    100.0|              2| 4428.0|Shipped|     4|      11|   2004|Classic Cars| 147|   S10_4962|Australia
n Collec...|   03 9520 4555|  636 St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|
Peter|  Medium|                     27|            100.0|       4428.0|
|      10370|             35|    65.63|              4|2297.05|Shipped|     1|       1|   2005|Classic Cars| 147|   S10_4962|Anna's De
coration...|   02 9936 8555|  201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|                     35|            65.63|      2297.05|
|      10391|             37|     46.9|              7| 1735.3|Shipped|     1|       3|   2005|Classic Cars| 147|   S10_4962|Anna's De
coration...|   02 9936 8555|  201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|   Small|                     37|             46.9|       1735.3|
|      10169|             30|    100.0|              2| 5019.9|Shipped|     4|      11|   2003|Classic Cars| 194|   S12_1099|Anna's De
coration...|   02 9936 8555|  201 Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|
Anna|  Medium|                     30|            100.0|       5019.9|
|      10127|             46|    100.0|              2|11279.2|Shipped|     2|       6|   2003|Classic Cars| 207|   S12_1108|   Muscle
Machine Inc|     2125557413|   4092 Furth Circle|   Suite 400|         NYC|      NY|     10022|      USA|       NA|          Young|
Jeff|   Large|                     46|            100.0|      11279.2|
+-----------+---------------+---------+---------------+-------+-------+------+--------+-------+------------+----+-----------+--------
-----------+---------------+--------------------+-------------+------------+--------+----------+---------+---------+---------------+---
-------------+--------+----------------------+-----------------+-------------+
only showing top 20 rows
```

[70]:

```python
imputed_df.head()
```

[70]:

```
Row(ORDERNUMBER=10223, QUANTITYORDERED=37, PRICEEACH=100.0, ORDERLINENUMBER=1, SALES=3965.66, STATUS='Shipped', QTR_ID=1, MONTH_ID=2,
YEAR_ID=2004, PRODUCTLINE='Motorcycles', MSRP=95, PRODUCTCODE='S10_1678', CUSTOMERNAME='Australian Collectors, Co.', PHONE='03 9520 45
55', ADDRESSLINE1='636 St Kilda Road', ADDRESSLINE2='Level 3', CITY='Melbourne', STATE='Victoria', POSTALCODE='3004', COUNTRY='Austral
ia', TERRITORY='APAC', CONTACTLASTNAME='Ferguson', CONTACTFIRSTNAME='Peter', DEALSIZE='Medium', QUANTITYORDERED_imputed=37, PRICEEACH_
imputed=100.0, SALES_imputed=3965.66)
```

[72]:

```python
imputed_df=imputed_df.drop('QUANTITYORDERED','PRICEEACH', 'SALES')
```

[73]:

```python
# Drop specified columns
columns_to_drop = ['QUANTITYORDERED','PRICEEACH', 'SALES']
imputed_df = imputed_df.drop(*columns_to_drop)

# Show the DataFrame after dropping columns
imputed_df.show()
```

```
--------------+-------------+-------------+---------+-----------+---------+---------+---------------+-----------------+---------+-----------
------------+------------------+-------------+
|ORDERNUMBER|ORDERLINENUMBER| STATUS|QTR_ID|MONTH_ID|YEAR_ID| PRODUCTLINE|MSRP|PRODUCTCODE|        CUSTOMERNAME|           PHONE|
ADDRESSLINE1|ADDRESSLINE2|        CITY|   STATE|POSTALCODE|  COUNTRY|TERRITORY|CONTACTLASTNAME|CONTACTFIRSTNAME|DEALSIZE|QUANTITYORDER
ED_imputed|PRICEEACH_imputed|SALES_imputed|
+-----------+---------------+-------+------+--------+-------+------------+----+-----------+--------------------+----------------+------
--------------+-------------+-------------+---------+-----------+---------+---------+---------------+-----------------+---------+-----------
------------+------------------+-------------+
|      10223|              1|Shipped|     1|       2|   2004| Motorcycles|  95|   S10_1678|Australian Collec...|    03 9520 4555|   636
St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|            Peter|  Medium|
37|            100.0|      3965.66|
|      10237|              7|Shipped|     2|       4|   2004| Motorcycles|  95|   S10_1678|     Vitachrome Inc.|      2125551500|   267
8 Kingston Rd.|   Suite 101|         NYC|      NY|     10022|      USA|       NA|          Frick|          Michael|   Small|
23|            100.0|      2333.12|
|      10361|             13|Shipped|     4|      12|   2004| Motorcycles|  95|   S10_1678|Souveniers And Th...|+61 2 9495 8555|Monito
r Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|           Adrian|   Small|
20|            72.55|       1451.0|
|      10163|              1|Shipped|     4|      10|   2003|Classic Cars| 214|   S10_1949|Classic Legends Inc.|      2125558493|    59
05 Pompton St.|   Suite 750|         NYC|      NY|     10022|      USA|       NA|      Hernandez|            Maria|  Medium|
21|            100.0|      4860.24|
|      10270|              9|Shipped|     3|       7|   2004|Classic Cars| 214|   S10_1949|Souveniers And Th...|+61 2 9495 8555|Monito
r Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|           Adrian|  Medium|
21|            100.0|      4905.39|
|      10347|              1|Shipped|     4|      11|   2004|Classic Cars| 214|   S10_1949|Australian Collec...|    03 9520 4555|   636
St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|            Peter|  Medium|
30|            100.0|       3944.7|
|      10391|              4|Shipped|     1|       3|   2005|Classic Cars| 214|   S10_1949|Anna's Decoration...|    02 9936 8555|   201
Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|             Anna|   Small|
24|            100.0|      2416.56|
|      10120|              3|Shipped|     2|       4|   2003| Motorcycles| 118|   S10_2016|Australian Collec...|    03 9520 4555|   636
St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|            Peter|   Small|
29|            96.34|      2793.86|
|      10223|              4|Shipped|     1|       2|   2004| Motorcycles| 118|   S10_2016|Australian Collec...|    03 9520 4555|   636
St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|            Peter|  Medium|
47|            100.0|      5422.39|
|      10361|              8|Shipped|     4|      12|   2004| Motorcycles| 118|   S10_2016|Souveniers And Th...|+61 2 9495 8555|Monito
r Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|           Adrian|   Small|
26|            51.15|       1329.9|
|      10120|              2|Shipped|     2|       4|   2003| Motorcycles| 193|   S10_4698|Australian Collec...|    03 9520 4555|   636
St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|            Peter|   Large|
46|            100.0|      9264.86|
|      10223|              3|Shipped|     1|       2|   2004| Motorcycles| 193|   S10_4698|Australian Collec...|    03 9520 4555|   636
St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|            Peter|   Large|
49|            100.0|      9774.03|
|      10237|              9|Shipped|     2|       4|   2004| Motorcycles| 193|   S10_4698|     Vitachrome Inc.|      2125551500|   267
8 Kingston Rd.|   Suite 101|         NYC|      NY|     10022|      USA|       NA|          Frick|          Michael|   Large|
39|            100.0|       7023.9|
|      10337|              8|Shipped|     4|      11|   2004|Classic Cars| 136|   S10_4757|Classic Legends Inc.|      2125558493|    59
05 Pompton St.|   Suite 750|         NYC|      NY|     10022|      USA|       NA|      Hernandez|            Maria|   Small|
25|            48.05|      1201.25|
|      10270|              2|Shipped|     3|       7|   2004|Classic Cars| 147|   S10_4962|Souveniers And Th...|+61 2 9495 8555|Monito
r Money Bui...|     Level 6|   Chatswood|     NSW|      2067|Australia|     APAC|         Huxley|           Adrian|  Medium|
32|            100.0|      4302.08|
|      10347|              2|Shipped|     4|      11|   2004|Classic Cars| 147|   S10_4962|Australian Collec...|    03 9520 4555|   636
St Kilda Road|     Level 3|   Melbourne|Victoria|      3004|Australia|     APAC|       Ferguson|            Peter|  Medium|
27|            100.0|       4428.0|
|      10370|              4|Shipped|     1|       1|   2005|Classic Cars| 147|   S10_4962|Anna's Decoration...|    02 9936 8555|   201
Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|             Anna|   Small|
35|            65.63|      2297.05|
|      10391|              7|Shipped|     1|       3|   2005|Classic Cars| 147|   S10_4962|Anna's Decoration...|    02 9936 8555|   201
Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|             Anna|   Small|
37|             46.9|       1735.3|
|      10169|              2|Shipped|     4|      11|   2003|Classic Cars| 194|   S12_1099|Anna's Decoration...|    02 9936 8555|   201
Miller Street|    Level 15|North Sydney|     NSW|      2060|Australia|     APAC|         O'Hara|             Anna|  Medium|
30|            100.0|       5019.9|
|      10127|              2|Shipped|     2|       6|   2003|Classic Cars| 207|   S12_1108| Muscle Machine Inc|      2125557413|   409
2 Furth Circle|   Suite 400|         NYC|      NY|     10022|      USA|       NA|          Young|             Jeff|   Large|
46|            100.0|      11279.2|
+-----------+---------------+-------+------+--------+-------+------------+----+-----------+--------------------+----------------+------
--------------+-------------+-------------+---------+-----------+---------+---------+---------------+-----------------+---------+-----------
------------+------------------+-------------+
only showing top 20 rows
```

[74]:

```
imputed_df.write.csv("imputed_data.csv", header=True)
```