```
pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.11/dist-packages (3.5.5)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.11/dist-packages (from pyspark) (0.10
```

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("Top Selling products").getOrCreate()
```

```
sc = spark.sparkContext
```

```
txnRDD = sc.textFile("/content/sample_data/txns1.txt")
```

```
txnRDD.count()
```

```
50000
```

```
txnRDD.getNumPartitions()
```

```
2
```

```
KVPairRDD = txnRDD.map(lambda a : (a.split(",")[5], float(a.split(",")[3])))
```

```
for line in KVPairRDD.take(5):
  print(line)
```

```
('Cardio Machine Accessories', 40.33)
('Weightlifting Gloves', 198.44)
('Weightlifting Machine Accessories', 5.58)
('Gymnastics Rings', 198.19)
('Field Hockey', 98.81)
```

```
SpentbyProd = KVPairRDD.reduceByKey(lambda a,b : a+b)
```

```
SpentbyProd.count()
```

```
125
```

```
for line in SpentbyProd.collect():
  print(line)
```

```
('Weightlifting Machines', 42701.40000000001)
```

```
('Ping Pong', 39973.020000000004)
('Outdoor Playsets', 39532.590000000026)
('Cardio Machine Accessories', 46485.54000000001)
('Weightlifting Gloves', 38438.720000000016)
('Weightlifting Machine Accessories', 41571.109999999986)
('Gymnastics Rings', 39871.54000000001)
('Camping & Backpacking & Hiking', 39993.52)
('Bungee Jumping', 38975.59)
('Archery', 37088.65999999999)
('Bowling', 40052.86000000001)
('Vaulting Horses', 41052.8)
('Baseball', 37843.82000000001)
('Weightlifting Belts', 45111.67999999999)
('Parachutes', 41186.419999999984)
('Kitesurfing', 37730.89)
('Mahjong', 44995.199999999975)
('Cricket', 37061.58000000001)
('Swimming', 43486.89000000003)
('Dice & Dice Sets', 41652.66000000002)
('Soccer', 39094.649999999994)
('Indoor Volleyball', 42146.44)
('Board Games', 41628.470000000016)
('Football', 42016.180000000015)
('Shooting Games', 41839.129999999976)
('Tetherball', 35611.92999999999)
('Water Polo', 43577.83)
('Exercise Bands', 37679.749999999985)
('Windsurfing', 43018.68000000001)
('Snowboarding', 38064.80999999999)
('Beach Volleyball', 44890.66999999999)
('Poker Chips & Sets', 42007.830000000016)
('Ballet Bars', 42603.71000000001)
('Softball', 40437.26000000001)
('Portable Electronic Games', 41931.249999999985)
('Trampolines', 42556.970000000016)
```

```
sortbyval = SpentbyProd.sortBy(lambda a : -a[1])
```

```
for line in sortbyval.collect():
  print(line)
```

```
('Yoga & Pilates', 47804.93999999999)
('Swing Sets', 47204.14)
('Lawn Games', 46828.43999999999)
('Golf', 46577.68)
('Cardio Machine Accessories', 46485.54000000001)
('Exercise Balls', 45143.84)
('Weightlifting Belts', 45111.67999999999)
('Mahjong', 44995.199999999975)
('Basketball', 44954.67999999999)
('Beach Volleyball', 44890.66999999999)
('Badminton', 44786.19000000001)
('Boxing', 44516.869999999995)
('Stopwatches', 44443.520000000004)
('Hockey', 44144.750000000015)
('Balance Beams', 44052.90000000001)
('Rugby', 43752.19000000002)
('Water Polo', 43577.83)
('Cross-Country Skiing', 43562.229999999996)
('Swimming', 43486.89000000003)
('Weight Benches', 43473.69000000001)
('Deck Shuffleboard', 43440.520000000004)
('Table Shuffleboard', 43405.15)
('Abdominal Equipment', 43304.109999999986)
('Darts', 43243.41999999998)
('Gymnastics Mats', 43224.55)
('Bobsledding', 43157.45999999996)
('Foosball', 43055.95999999999)
('Boating', 43049.06999999999)
('Windsurfing', 43018.68000000001)
('Medicine Balls', 42798.859999999986)
('Foam Rollers', 42779.17999999999)
('Lacrosse', 42732.61)
('Trampoline Accessories', 42726.340000000026)
```

```
('Weightlifting Machines', 42701.40000000001)
('Skateboarding', 42632.18)
('Ballet Bars', 42603.71000000001)
('Trampolines', 42556.970000000016)
('Sandboxes', 42535.79999999999)
('Bodyboarding', 42457.489999999976)
('Skating', 42443.56000000001)
('Motorsports', 42427.04999999998)
('Cycling', 42243.91)
('Playhouses', 42186.76999999999)
('Racquetball', 42183.48999999999)
('Water Tubing', 42154.94)
('Indoor Volleyball', 42146.44)
('Football', 42016.180000000015)
('Poker Chips & Sets', 42007.830000000016)
('Free Weights', 41966.600000000006)
('Portable Electronic Games', 41931.249999999985)
('Free Weight Bars', 41915.61999999999)
('Bingo Sets', 41896.56000000001)
('Shooting Games', 41839.129999999976)
('Lawn Water Slides', 41730.19000000002)
('Dice & Dice Sets', 41652.66000000002)
('Board Games', 41628.470000000016)
('Weightlifting Machine Accessories', 41571.109999999986)
('Cheerleading', 41244.57000000001)
```

```
csv_formatRDD = sortbyval.map(lambda a : (a[0] + "," + str(round(a[1],2))  ))
```

```
for line in csv_formatRDD.take(5):
  print(line)
```

```
Yoga & Pilates,47804.94
Swing Sets,47204.14
Lawn Games,46828.44
Golf,46577.68
Cardio Machine Accessories,46485.54
```

```
csv_formatRDD.saveAsTextFile("/content/sample_data/spark3")
```

```
csv_formatRDD = csv_formatRDD.repartition(1)
```

```
csv_formatRDD.saveAsTextFile("/content/sample_data/spark4")
```