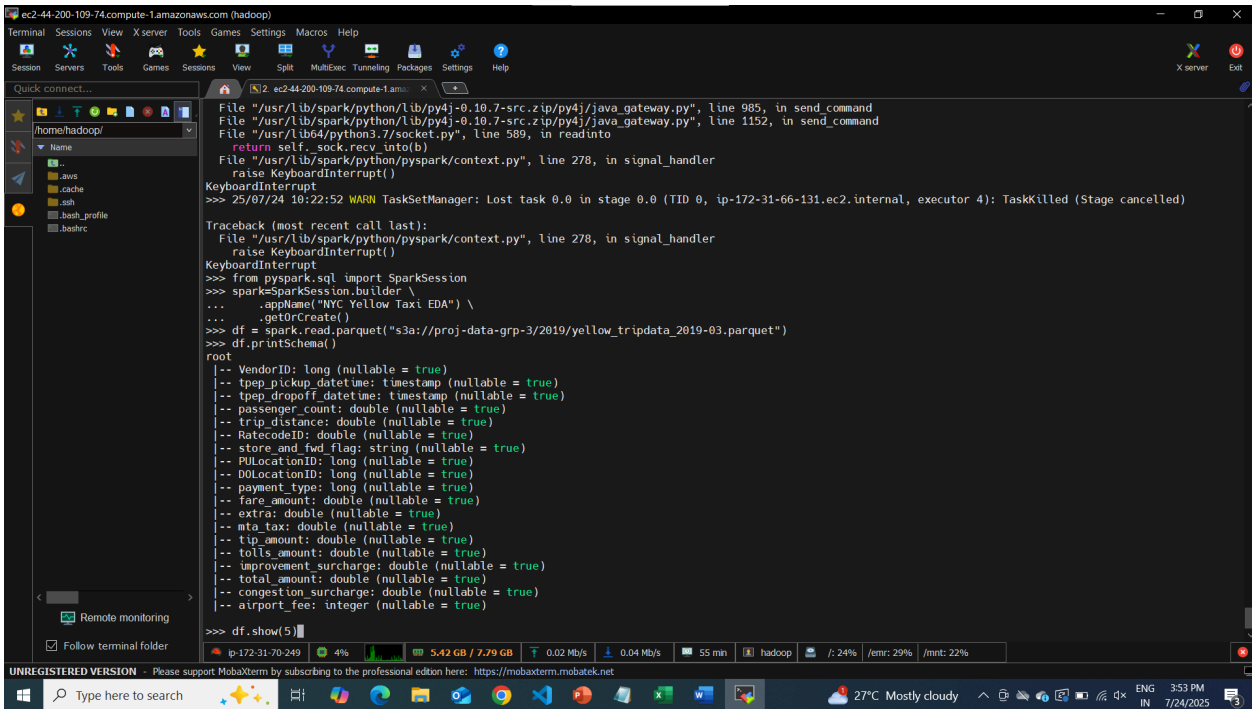
1.




```
>>> df.show(5)
+--------+--------------------+---------------------+---------------+-------------+----------+------------------+------------+------------+------------+------
------+-----+-------+----------+-----------+---------------------+-----------+--------------------+-----------+
|VendorID|tpep_pickup_datetime|tpep_dropoff_datetime|passenger_count|trip_distance|RatecodeID|store_and_fwd_flag|PULocationID|DOLocationID|payment_type|fare_a
mount|extra|mta_tax|tip_amount|tolls_amount|improvement_surcharge|total_amount|congestion_surcharge|airport_fee|
+--------+--------------------+---------------------+---------------+-------------+----------+------------------+------------+------------+------------+------
-----+-----+-------+----------+-----------+---------------------+-----------+--------------------+-----------+
|       1| 2019-03-01 00:24:41|  2019-03-01 00:25:31|            1.0|          0.0|       1.0|                 N|         145|         145|           2|
  2.5|  0.5|    0.5|       0.0|        0.0|                  0.3|        3.8|                 0.0|       null|
|       1| 2019-03-01 00:25:27|  2019-03-01 00:36:37|            2.0|          3.7|       1.0|                 N|          95|         130|           1|
 13.0|  0.5|    0.5|       0.7|        0.0|                  0.3|       15.0|                 0.0|       null|
|       1| 2019-03-01 00:05:21|  2019-03-01 00:38:23|            1.0|         14.1|       1.0|                 N|         249|          28|           1|
 41.0|  3.0|    0.5|      10.1|       5.76|                  0.3|      60.66|                 2.5|       null|
|       1| 2019-03-01 00:48:55|  2019-03-01 01:06:03|            1.0|          9.6|       1.0|                 N|         138|          98|           2|
 27.0|  0.5|    0.5|       0.0|        0.0|                  0.3|       28.3|                 0.0|       null|
|       1| 2019-03-01 00:11:42|  2019-03-01 00:16:40|            1.0|          0.8|       1.0|                 N|          48|          48|           1|
  5.5|  3.0|    0.5|       3.0|        0.0|                  0.3|       12.3|                 2.5|       null|
+--------+--------------------+---------------------+---------------+-------------+----------+------------------+------------+------------+------------+------
-----+-----+-------+----------+-----------+---------------------+-----------+--------------------+-----------+
only showing top 5 rows

>>>
```

```
>>> # Columns and types
>>> df.printSchema()
root
 |-- VendorID: long (nullable = true)
 |-- tpep_pickup_datetime: timestamp (nullable = true)
 |-- tpep_dropoff_datetime: timestamp (nullable = true)
 |-- passenger_count: double (nullable = true)
 |-- trip_distance: double (nullable = true)
 |-- RatecodeID: double (nullable = true)
 |-- store_and_fwd_flag: string (nullable = true)
 |-- PULocationID: long (nullable = true)
 |-- DOLocationID: long (nullable = true)
 |-- payment_type: long (nullable = true)
 |-- fare_amount: double (nullable = true)
 |-- extra: double (nullable = true)
 |-- mta_tax: double (nullable = true)
 |-- tip_amount: double (nullable = true)
 |-- tolls_amount: double (nullable = true)
 |-- improvement_surcharge: double (nullable = true)
 |-- total_amount: double (nullable = true)
 |-- congestion_surcharge: double (nullable = true)
 |-- airport_fee: integer (nullable = true)

>>>
```

```
>>> # Summary stats
>>> df.describe().show()
+-------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+
|summary|          VendorID|   passenger_count|     trip_distance|        RatecodeID|store_and_fwd_flag|      PULocationID|      DOLocationID|      payment_type|       fare_amount|             extra|           mta_tax|        tip_amount|      tolls_amount|improvement_surcharge|      total_amount|congestion_surcharge|       airport_fee|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+---------------------+------------------+--------------------+------------------+
|  count|           7866620|           7833146|           7866620|           7833146|           7833146|           7866620|           7866620|           7866620|           7866620|           7866620|           7866620|           7866620|           7866620|              7866620|           7866620|             7866620|           7866620|
|   mean|1.6378870467875657|1.5751072429902366| 3.023191633764328|1.0668542371098406|              null|161.88302561964352|160.15224759807896|1.2728699238046328|13.303418062653337|1.2077227081008084|0.4949409733786556|2.2224419954258554|0.37197059982261993|   0.2989747947748331| 19.31109651005561|  2.2811748306491415|              null|
| stddev|0.5128753403408546| 1.228311092883687|3.9280515814262924|0.6982221180871948|              null| 65.93281903603491|  70.2810794443417|0.47762656308391736|346.08384130635153|1.2865501865236388|0.07108076737503598| 50.52138315426142| 1.6870989308702269|  0.024010033128652744|395.37424676447057|  0.7145033289268481|              null|
|    min|                 1|               0.0|               0.0|               1.0|                 N|                 1|                 1|                 0|            -447.0|             -57.5|              -0.5|             -89.89|             -25.0|                 -0.3|            -450.3|                -2.5|              null|
|    max|                 5|               9.0|            237.17|             99.0|                 Y|               265|               265|                 5|          943274.8|             24.66|              75.0|         141492.02|             805.76|                  1.0|         1084772.17|                 4.5|              null|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+------------------+---------------------+------------------+--------------------+------------------+
```

3. **Initial Cleaning & Understanding**

```
>>> # Null count per column
>>> from pyspark.sql.functions import col, isnan, when, count
>>> df.select([count(when(col(c).isNull(), c)).alias(c) for c in df.columns]).show()
+--------+--------------------+---------------------+---------------+-------------+------------+----------+------------------+------------+------------+------------+------------+-----------+
|VendorID|tpep_pickup_datetime|tpep_dropoff_datetime|passenger_count|trip_distance|RatecodeID|store_and_fwd_flag|PULocationID|DOLocationID|payment_type|fare_amount|extra|mta_tax|tip_amount|tolls_amount|improvement_surcharge|total_amount|congestion_surcharge|airport_fee|
+--------+--------------------+---------------------+---------------+-------------+----------+------------------+------------+------------+------------+-----------+-----+-------+----------+------------+---------------------+------------+--------------------+-----------+
|       0|                   0|                    0|          33474|            0|     33474|             33474|           0|           0|           0|          0|    0|      0|         0|           0|                    0|     7866620|                   0|          0|
+--------+--------------------+---------------------+---------------+-------------+----------+------------------+------------+------------+------------+-----------+-----+-------+----------+------------+---------------------+------------+--------------------+-----------+
```

1. **Trip Count per Month**

```
>>> from pyspark.sql.functions import month
>>>
>>> df.withColumn("month", month("tpep_pickup_datetime")) \
...   .groupBy("month") \
...   .count() \
...   .orderBy("month") \
...   .show()
+-----+-------+
|month|  count|
+-----+-------+
|    1|     42|
|    2|    362|
|    3|7866111|
|    4|     78|
|    5|     11|
|    6|      1|
|    7|      5|
|    8|      1|
|   12|      9|
+-----+-------+
```

2. **Trip Distance Distribution**

```
>>> df.select("trip_distance").describe().show()
+-------+------------------+
|summary|     trip_distance|
+-------+------------------+
|  count|           7866620|
|   mean| 3.023191633764328|
| stddev|3.9280515814262924|
|    min|               0.0|
|    max|            237.17|
+-------+------------------+
```

3. **Fare vs. Distance Relationship**

```
>>> df.select("fare_amount", "trip_distance") \
...    .filter("trip_distance > 0 and fare_amount > 0") \
...    .show(10)
+-----------+-------------+
|fare_amount|trip_distance|
+-----------+-------------+
|       13.0|          3.7|
|       41.0|         14.1|
|       27.0|          9.6|
|        5.5|          0.8|
|        6.0|          1.2|
|        5.5|          0.6|
|       17.0|         5.65|
|        6.0|         1.16|
|        5.0|         0.71|
|       10.5|         2.63|
+-----------+-------------+
only showing top 10 rows
```

4. **Top Pickup & Drop-off Locations**

```
>>> # Top Pickup Locations
>>> df.groupBy("PULocationID") \
...    .count() \
...    .orderBy(col("count").desc()) \
...    .show(10)
[Stage 17:=============================================>            (3 + 1) / 4]25/07/24 10:46:23 WARN Utils: Truncated the string representation
nce it was too large. This behavior can be adjusted by setting 'spark.debug.maxToStringFields' in SparkEnv.conf.
+------------+------+
|PULocationID| count|
+------------+------+
|         161|328433|
|         237|311896|
|         162|286072|
|         236|283119|
|         230|279261|
|         186|278438|
|          48|256545|
|         234|248504|
|         170|243858|
|         142|229982|
+------------+------+
only showing top 10 rows
```

```
>>>
>>> # Top Drop-off Locations
>>> df.groupBy("DOLocationID") \
...   .count() \
...   .orderBy(col("count").desc()) \
...   .show(10)
+------------+------+
|DOLocationID| count|
+------------+------+
|         161|300488|
|         236|300236|
|         237|283692|
|         170|250462|
|         230|238802|
|         162|233992|
|          48|223673|
|         234|211565|
|         142|208331|
|         186|197788|
+------------+------+
only showing top 10 rows
```

5. **Average Trip Duration**

```
>>> from pyspark.sql.functions import unix_timestamp
>>> df.withColumn("trip_duration_min",
...               (unix_timestamp("tpep_dropoff_datetime") - unix_timestamp("tpep_pickup_datetime")) / 60) \
...   .select("trip_duration_min") \
...   .describe() \
...   .show()
+-------+------------------+
|summary| trip_duration_min|
+-------+------------------+
|  count|           7866620|
|   mean|17.729944578056063|
| stddev| 71.28184046473208|
|    min|-4139.016666666666|
|    max|27454.783333333333|
+-------+------------------+
```

6. **Payment Type Distribution**

```
>>> df.groupBy("payment_type") \
...   .count() \
...   .orderBy(col("count").desc()) \
...   .show()
+------------+-------+
|payment_type|  count|
+------------+-------+
|           1|5721775|
|           2|2057412|
|           3|  39281|
|           0|  33474|
|           4|  14648|
|           5|     30|
+------------+-------+

>>>
```

7. **Tip Analysis**

```
>>> df.select("tip_amount").describe().show()
+-------+------------------+
|summary|        tip_amount|
+-------+------------------+
|  count|           7866620|
|   mean|2.2224419954258554|
| stddev| 50.52138315426142|
|    min|            -89.89|
|    max|         141492.02|
+-------+------------------+

>>>
>>> # Tip-to-Fare Ratio (optional)
>>> df.withColumn("tip_ratio", col("tip_amount") / col("fare_amount")) \
...   .select("tip_ratio") \
...   .describe() \
...   .show()
+-------+-------------------+
|summary|          tip_ratio|
+-------+-------------------+
|  count|            7863649|
|   mean|0.19918917663286123|
| stddev| 11.090063586122845|
|    min|-0.4523076923076923|
|    max|            22777.0|
+-------+-------------------+
```

8. **Passenger Count Distribution**

```
>>> df.groupBy("passenger_count") \
...    .count() \
...    .orderBy("passenger_count") \
...    .show()
+---------------+-------+
|passenger_count|  count|
+---------------+-------+
|           null|  33474|
|            0.0| 140131|
|            1.0|5496239|
|            2.0|1176562|
|            3.0| 329128|
|            4.0| 152587|
|            5.0| 340192|
|            6.0| 198225|
|            7.0|     40|
|            8.0|     18|
|            9.0|     24|
+---------------+-------+
```

9. **Missing Value Check**

```
>>> from pyspark.sql.functions import col, sum as spark_sum
>>>
>>> df.select([spark_sum(col(c).isNull().cast("int")).alias(c) for c in df.columns]).show()
+--------+--------------------+---------------------+---------------+-------------+----------+------------------+------------+------------+------------+------
------+-----+-------+----------+-----------+--------------------+-----------+--------------------+-----------+
|VendorID|tpep_pickup_datetime|tpep_dropoff_datetime|passenger_count|trip_distance|RatecodeID|store_and_fwd_flag|PULocationID|DOLocationID|payment_type|fare_a
mount|extra|mta_tax|tip_amount|tolls_amount|improvement_surcharge|total_amount|congestion_surcharge|airport_fee|
+--------+--------------------+---------------------+---------------+-------------+----------+------------------+------------+------------+------------+------
------+-----+-------+----------+-----------+--------------------+-----------+--------------------+-----------+
|       0|                   0|                    0|          33474|            0|     33474|             33474|           0|           0|           0|
    0|    0|      0|         0|          0|                   0|          0|               33474|    7866620|
+--------+--------------------+---------------------+---------------+-------------+----------+------------------+------------+------------+------------+------
------+-----+-------+----------+-----------+--------------------+-----------+--------------------+-----------+

>>>
```

10. **Trip Duration Outliers**

```
>>> # Duration > 6 hours
>>> df.filter("trip_duration_min > 360").show()
+--------+--------------------+--------------------+---------------+-------------+----------+------------------+------------+------------+------------+-----+-----+-------+----------+-----------+---------------------+------------+--------------------+----------+-----------------+
|VendorID|tpep_pickup_datetime|tpep_dropoff_datetime|passenger_count|trip_distance|RatecodeID|store_and_fwd_flag|PULocationID|DOLocationID|payment_type|fare_amount|extra|mta_tax|tip_amount|tolls_amount|improvement_surcharge|total_amount|congestion_surcharge|airport_fee| trip_duration_min|
+--------+--------------------+--------------------+---------------+-------------+----------+------------------+------------+------------+------------+-----+-----+-------+----------+-----------+---------------------+------------+--------------------+----------+-----------------+
|       2| 2019-02-28 21:34:06| 2019-03-01 20:41:16|            1.0|         1.79|       1.0|                 N|          65|          61|           2|  7.5|  0.5|    0.5|       0.0|        0.0|                  0.3|         8.8|                 0.0|      null|1387.1666666666667|
|       2| 2019-03-01 00:49:35| 2019-03-02 00:27:36|            2.0|         0.35|       1.0|                 N|          79|          79|           2|  5.5|  0.5|    0.5|       0.0|        0.0|                  0.3|         9.3|                 2.5|      null|1418.0166666666667|
|       2| 2019-03-01 00:32:11| 2019-03-01 23:45:36|            2.0|         2.44|       1.0|                 N|          79|         246|           1| 11.0|  0.5|    0.5|      2.96|        0.0|                  0.3|       17.76|                 2.5|      null|1393.4166666666667|
|       2| 2019-02-28 23:58:18| 2019-03-01 23:56:12|            2.0|         6.23|       1.0|                 N|         246|         255|           2| 30.5|  0.5|    0.5|       0.0|        0.0|                  0.3|        34.3|                 2.5|      null|            1437.9|
|       2| 2019-02-28 09:38:07| 2019-03-01 09:04:26|            1.0|         4.33|       1.0|                 N|          90|         256|           1| 19.5|  0.5|    0.5|      4.66|        0.0|                  0.3|       27.96|                 2.5|      null|1406.3166666666666|
|       2| 2019-02-28 22:32:43| 2019-03-01 21:36:17|            2.0|         0.78|       1.0|                 N|         162|         233|           1|  4.5|  0.5|    0.5|      2.49|        0.0|                  0.3|       10.79|                 2.5|      null|1383.5666666666666|
|       2| 2019-02-28 16:30:50| 2019-03-01 15:38:48|            1.0|          1.5|       1.0|                 N|         244|         244|           2|  8.0|  0.5|    0.5|       0.0|        0.0|                  0.3|         9.3|                 0.0|      null|1387.9666666666667|
|       2| 2019-03-01 00:01:32| 2019-03-01 23:59:18|            5.0|         5.93|       1.0|                 N|         229|          65|           1| 30.5|  0.5|    0.5|       0.0|        0.0|                  0.3|        34.3|                 2.5|      null|1437.7666666666667|
|       2| 2019-02-28 15:04:48| 2019-03-01 14:18:50|            1.0|         1.71|       1.0|                 N|         229|          48|           1| 10.0|  0.5|    0.5|      3.45|        0.0|                  0.3|       17.25|                 2.5|      null|1394.0333333333333|
|       2| 2019-03-01 00:48:09| 2019-03-02 00:04:31|            5.0|         8.67|       1.0|                 N|         138|          17|           1| 25.0|  0.5|    0.5|      5.26|        0.0|                  0.3|       31.56|                 0.0|      null|1396.3666666666666|
|       2| 2019-03-01 00:28:59| 2019-03-02 00:25:10|            3.0|         0.88|       1.0|                 N|         246|          68|           1|  5.5|  0.5|    0.5|      0.93|        0.0|                  0.3|       10.23|                 2.5|      null|1436.1833333333334|
|       2| 2019-02-28 16:37:05| 2019-03-01 16:29:19|            2.0|         5.32|       1.0|                 N|         164|          13|           1| 19.0|  0.5|    0.5|       0.0|        0.0|                  0.3|        22.8|                 2.5|      null|1432.2333333333333|
|       2| 2019-03-01 00:42:38| 2019-03-02 00:40:12|            2.0|         1.03|       1.0|                 N|         166|         238|           2|  6.0|  0.5|    0.5|       0.0|        0.0|                  0.3|         7.3|                 0.0|      null|1437.5666666666666|
|       2| 2019-03-01 00:56:51| 2019-03-02 00:55:01|            5.0|         0.73|       1.0|                 N|         230|         186|           2|  5.0|  0.5|    0.5|       0.0|        0.0|                  0.3|         8.8|                 2.5|      null|1438.1666666666667|
|       2| 2019-02-28 02:16:08| 2019-03-01 01:39:14|            1.0|         4.63|       1.0|                 N|          79|         142|           1| 18.0|  0.5|    0.5|      4.36|        0.0|                  0.3|       26.16|                 2.5|      null|            1403.1|
```

```
| 18.5|  0.5|    0.5|      2.23|        0.0|                  0.3|       24.53|                 2.5|      null|            1428.05|
|       2| 2019-02-28 14:49:42| 2019-03-01 14:21:29|            1.0|         7.23|       1.0|                 N|          90|          61|           1|
| 27.0|  0.5|    0.5|       7.7|        0.0|                  0.3|        38.5|                 2.5|      null|1411.7833333333333|
|       2| 2019-03-01 00:55:02| 2019-03-02 00:26:07|            1.0|         1.91|       1.0|                 N|         107|         211|           1|
| 10.0|  0.5|    0.5|      2.76|        0.0|                  0.3|       16.56|                 2.5|      null|1411.0833333333333|
|       2| 2019-03-01 01:37:37| 2019-03-02 01:14:15|            2.0|         8.95|       1.0|                 N|         145|          33|           1|
| 29.0|  0.5|    0.5|       0.0|        0.0|                  0.3|        32.8|                 2.5|      null|1416.6333333333334|
+--------+--------------------+--------------------+---------------+-------------+----------+------------------+------------+------------+------------+-----+-----+-------+----------+-----------+---------------------+------------+--------------------+----------+-----------------+
```

```
>>> # Distance = 0 and Duration > 10 minutes
>>> df.filter((col("trip_distance") == 0) & (col("trip_duration_min") > 10)).show()
+--------+--------------------+--------------------+---------------+-------------+----------+------------------+------------+------------+------------+-----+-----+-------+----------+-----------+---------------------+------------+--------------------+----------+-----------------+
|VendorID|tpep_pickup_datetime|tpep_dropoff_datetime|passenger_count|trip_distance|RatecodeID|store_and_fwd_flag|PULocationID|DOLocationID|payment_type|fare_amount|extra|mta_tax|tip_amount|tolls_amount|improvement_surcharge|total_amount|congestion_surcharge|airport_fee| trip_duration_min|
+--------+--------------------+--------------------+---------------+-------------+----------+------------------+------------+------------+------------+-----+-----+-------+----------+-----------+---------------------+------------+--------------------+----------+-----------------+
|       2| 2019-03-01 00:25:59| 2019-03-01 01:19:28|            2.0|          0.0|       1.0|                 N|          17|          17|           2|  0.0|  0.0|    0.5|       0.0|        0.0|                  0.3|         0.8|                 0.0|      null|53.483333333333334|
|       1| 2019-03-01 00:30:11| 2019-03-01 00:50:14|            1.0|          0.0|       1.0|                 N|          79|          87|           2| 14.0|  3.0|    0.5|       0.0|        0.0|                  0.3|        17.8|                 2.5|      null|             20.05|
|       2| 2019-03-01 00:28:33| 2019-03-01 00:40:05|            4.0|          0.0|       1.0|                 N|          36|          17|           1|  2.5|  0.5|    0.5|      10.0|        0.0|                  0.3|        13.8|                 0.0|      null|11.533333333333333|
|       2| 2019-03-01 21:56:09| 2019-03-02 00:04:53|            1.0|          0.0|       1.0|                 N|         264|         264|           1|  0.0|  0.0|    0.0|       0.0|        0.0|                  0.0|         0.0|                 0.0|      null|128.73333333333332|
|       1| 2019-03-01 01:02:40| 2019-03-01 01:24:02|            1.0|          0.0|       1.0|                 N|          79|          37|           1| 17.0|  3.0|    0.5|       6.2|        0.0|                  0.3|        27.0|                 2.5|      null|21.366666666666667|
|       2| 2019-03-01 02:52:36| 2019-03-01 03:28:02|            1.0|          0.0|       5.0|                 N|         265|         265|           1| 94.0|  0.0|    0.0|     18.86|        0.0|                  0.3|      113.16|                 0.0|      null| 35.43333333333333|
|       1| 2019-03-01 03:05:48| 2019-03-01 03:19:23|            1.0|          0.0|       1.0|                 N|         231|         257|           1| 19.0|  3.0|    0.5|      4.55|        0.0|                  0.3|       27.35|                 2.5|      null|13.583333333333334|
|       2| 2019-03-01 03:59:29| 2019-03-01 04:11:19|            1.0|          0.0|       5.0|                 N|         264|         142|           1|  8.3|  0.0|    0.5|      2.73|        0.0|                  0.3|       11.83|                 0.0|      null|11.833333333333334|
|       2| 2019-03-01 03:35:24| 2019-03-01 04:16:28|            1.0|          0.0|       5.0|                 N|          68|         238|           1| 20.0|  0.0|    0.0|       4.5|        0.0|                  0.0|        27.0|                 2.5|      null| 41.06666666666667|
|       1| 2019-03-01 05:56:12| 2019-03-01 06:21:54|            1.0|          0.0|       1.0|                 N|         139|          93|           1| 74.4|  0.0|    0.5|       0.0|        0.0|                  0.3|        75.2|                 0.0|      null|             25.7|
|       1| 2019-03-01 05:48:48| 2019-03-01 06:21:24|            1.0|          0.0|       1.0|                 N|          69|          97|           1| 70.4|  0.0|    0.5|       0.0|        0.0|                  0.3|        71.2|                 0.0|      null|             32.6|
|       1| 2019-03-01 06:54:48| 2019-03-01 07:05:28|            1.0|          0.0|       1.0|                 N|         151|          74|           2|  9.5|  0.0|    0.5|       0.0|        0.0|                  0.3|        10.3|                 0.0|      null|10.666666666666666|
|       1| 2019-03-01 06:53:50| 2019-03-01 08:18:09|            1.0|          0.0|       1.0|                 Y|          16|          14|           1| 74.4|  0.0|    0.5|       0.0|        0.0|                  0.3|        75.2|                 0.0|      null| 84.31666666666666|
|       2| 2019-03-01 06:08:53| 2019-03-01 06:36:06|            0.0|          0.0|       5.0|                 N|         265|         265|           1|172.5|  0.0|    0.5|      22.0|        0.0|                  0.3|       195.3|                 0.0|      null|27.216666666666665|
|       1| 2019-03-01 07:19:43| 2019-03-01 07:30:38|            1.0|          0.0|       1.0|                 N|          43|         164|           1| 10.5|  2.5|    0.5|      2.75|        0.0|                  0.3|       16.55|                 2.5|      null|10.916666666666666|
|       2| 2019-03-01 06:28:52| 2019-03-01 07:28:06|            1.0|          0.0|       5.0|                 N|         186|         132|           1|
```

```
172.5|  0.0|    0.5|      22.0|          0.0|                0.3|       195.3|                0.0|       null|27.216666666666665|
|       1| 2019-03-01 07:19:43|  2019-03-01 07:30:38|           1.0|         0.0|       1.0|                  N|          43|        164|          1|
 10.5|  2.5|    0.5|      2.75|          0.0|                0.3|       16.55|                2.5|       null|10.916666666666666|
|       2| 2019-03-01 06:28:52|  2019-03-01 07:28:06|           1.0|         0.0|       5.0|                  N|         186|        132|          1|
 61.0|  0.0|    0.0|     19.05|          0.0|                0.0|       82.55|                2.5|       null|59.233333333333334|
|       1| 2019-03-01 07:08:53|  2019-03-01 08:26:39|           1.0|         0.0|       1.0|                  N|         219|        231|          1|
 74.4|  0.0|    0.5|       0.0|          0.0|                0.3|        75.2|                0.0|       null| 77.76666666666667|
|       2| 2019-03-01 08:15:14|  2019-03-01 08:47:36|           1.0|         0.0|       1.0|                  N|          75|        239|          2|
  2.5|  0.0|    0.5|       0.0|          0.0|                0.3|         5.8|                2.5|       null| 32.36666666666667|
|       1| 2019-03-01 08:49:40|  2019-03-01 09:59:10|           1.0|         0.0|       1.0|                  N|          72|        244|          1|
 70.4|  0.0|    0.5|       0.0|         5.76|                0.3|       76.96|                0.0|       null|              69.5|
|       2| 2019-03-01 07:57:44|  2019-03-01 08:19:31|           2.0|         0.0|       1.0|                  N|          85|         89|          3|
 -2.5|  0.0|   -0.5|       0.0|          0.0|               -0.3|        -3.3|                0.0|       null|21.783333333333335|
+--------+--------------------+---------------------+--------------+------------+----------+------------------+------------+------------+-----
----+-----+-------+----------+------------+---------------------+------------+------------------+----------+------------------+
only showing top 20 rows
```

6. **Example – Trip Duration Calculation**

```
>>> from pyspark.sql.functions import unix_timestamp, col
>>>
>>> df = df.withColumn("trip_duration_min",
...                    (unix_timestamp("tpep_dropoff_datetime") - unix_timestamp("tpep_pickup_datetime")) / 60)
>>>
>>> df.select("trip_duration_min").describe().show()
+-------+------------------+
|summary|  trip_duration_min|
+-------+------------------+
|  count|           7866620|
|   mean|17.729944578056063|
| stddev| 71.28184046473208|
|    min|-4139.016666666666|
|    max|27454.783333333333|
+-------+------------------+
```

7. **Convert to Pandas for Plotting (optional)**

8. Save Cleaned/Analyzed Data to S3 (optional)