```
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.11/dist-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.11/dist-packages (from pyspark) (0.10.9.7)
```

```
from pyspark.sql import SparkSession

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

```
df = spark.read.parquet("/content/yellow_tripdata_2020-01.parquet")
df.show(5)
```

```
+--------+--------------------+--------------------+---------------+-------------+----------+------------------+---------
|VendorID|tpep_pickup_datetime|tpep_dropoff_datetime|passenger_count|trip_distance|RatecodeID|store_and_fwd_flag|PULocatic
+--------+--------------------+--------------------+---------------+-------------+----------+------------------+---------
|       1| 2020-01-01 00:28:15|  2020-01-01 00:33:03|            1.0|          1.2|       1.0|                 N|
|       1| 2020-01-01 00:35:39|  2020-01-01 00:43:04|            1.0|          1.2|       1.0|                 N|
|       1| 2020-01-01 00:47:41|  2020-01-01 00:53:52|            1.0|          0.6|       1.0|                 N|
|       1| 2020-01-01 00:55:23|  2020-01-01 01:00:14|            1.0|          0.8|       1.0|                 N|
|       2| 2020-01-01 00:01:58|  2020-01-01 00:04:16|            1.0|          0.0|       1.0|                 N|
+--------+--------------------+--------------------+---------------+-------------+----------+------------------+---------
only showing top 5 rows
```

```
df.printSchema()
```

```
root
 |-- VendorID: long (nullable = true)
 |-- tpep_pickup_datetime: timestamp_ntz (nullable = true)
 |-- tpep_dropoff_datetime: timestamp_ntz (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)
```

```
df.describe().show()
```

```
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
|summary|          VendorID|   passenger_count|    trip_distance|        RatecodeID|store_and_fwd_flag|      PULocationID|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
|  count|           6405008|           6339567|           6405008|           6339567|           6339567|           6405008|
|   mean|1.6730021258365328|1.5153326717739555| 2.92964393330939|1.0599077192495954|              NULL|164.73225778952968|
| stddev|0.4691264930553467|1.1515942134278447|83.15910597325033|0.8118432071906562|              NULL| 65.54373944111883|
|    min|                 1|                 0|               0.0|               1.0|                 N|                 1|
|    max|                 2|                 9|          66031.0|              99.0|                 Y|               265|
+-------+------------------+------------------+------------------+------------------+------------------+------------------+
```

```
total_trips = df.count()
print(f"Total trips in dataset: {total_trips}")
```

```
Total trips in dataset: 6405008
```

```
from pyspark.sql.functions import unix_timestamp, col

df = df.withColumn("trip_duration_minutes",
    (unix_timestamp("tpep_dropoff_datetime") - unix_timestamp("tpep_pickup_datetime")) / 60)

df.select("tpep_pickup_datetime", "tpep_dropoff_datetime", "trip_duration_minutes").show(5)
```

```
+--------------------+---------------------+--------------------+
|tpep_pickup_datetime|tpep_dropoff_datetime|trip_duration_minutes|
+--------------------+---------------------+--------------------+
| 2020-01-01 00:28:15|  2020-01-01 00:33:03|                 4.8|
| 2020-01-01 00:35:39|  2020-01-01 00:43:04|   7.416666666666667|
| 2020-01-01 00:47:41|  2020-01-01 00:53:52|   6.183333333333334|
| 2020-01-01 00:55:23|  2020-01-01 01:00:14|                4.85|
| 2020-01-01 00:01:58|  2020-01-01 00:04:16|                 2.3|
+--------------------+---------------------+--------------------+
only showing top 5 rows
```

```
df.selectExpr("avg(fare_amount) as avg_fare").show()
```

```
+-----------------+
|         avg_fare|
+-----------------+
|12.69410811978051|
+-----------------+
```

```
df.groupBy("payment_type").count().show()
```

```
+------------+-------+
|payment_type|  count|
+------------+-------+
|           0|  65441|
|           5|      1|
|           1|4694897|
|           3|  32770|
|           2|1593834|
|           4|  18065|
+------------+-------+
```

```
df.write.csv("/content/cleaned_nyc_taxi.csv", header=True)
```