```
Line 1
from pyspark.sql import SparkSession
from pyspark.sql import Row
```

```
Line 2
spark = SparkSession.builder.appName("SparkSQL").getOrCreate()
```

```
Line 3
df = spark.read.option("header", "true").option("inferSchema", "true")\
    .csv("fhv_tripdata_2019-01.csv")
print("Here is our inferred schema:")
df.printSchema()
```

```
Line 4
df.show()
```

```
Line 5
df.createOrReplaceTempView("tripData")
```

```
Line 6
sqlDF = spark.sql("SELECT DISTINCT dispatching_base_num,  pickup_datetime,
dropoff_datetime FROM \
tripData where PULocationID is not null and DOLocationID ='264' and
dispatching_base_num= 'B02182' \
GROUP BY dispatching_base_num,  pickup_datetime, dropoff_datetime ORDER BY
pickup_datetime ASC")
sqlDF.show()
```

```
Line 7
df_distinct = spark.sql("SELECT DISTINCT SR_Flag from tripData where PULocationID
IS NOT NULL and DOLocationID is NOT NULL ")
df_distinct.show()
```

```
Line 8
sqlDF.write.parquet("spark_write_parquet.parquet")
```

```
Line 9
df_parquet = spark.read.parquet("spark_write_parquet.parquet")
df_parquet.show()
```

```
Line 10
sqlTripData = spark.sql("SELECT dispatching_base_num, PUlocationID, DOlocationID,
Affiliated_base_number \
FROM tripData WHERE PULocationID IS NOT NULL and DOLocationID is NOT NULL AND
pickup_datetime >= '2019-01-01' \
AND pickup_datetime <='2019-01-10' GROUP BY dispatching_base_num, PUlocationID,
DOlocationID, Affiliated_base_number \
ORDER BY dispatching_base_num")
sqlTripData.show()
```

You can choose **fhv_tripdata_2019-01.csv.gz** and download it

https://github.com/DataTalksClub/nyc-tlc-data/releases/tag/fhv