dsti-a19/ajourdan/lab5

FINISHED

```
from pyspark.sql import functions as F

# Define the schema of the nycTaxiFares dataset
fares_schema = StructType([
    StructField('ride_id', IntegerType(), False),
    StructField('taxi_id', IntegerType(), False),
    StructField('driver_id', IntegerType(), False),
    StructField('start_time', TimestampType(), False),
    StructField('payment_type', StringType(), False),
    StructField('tip', FloatType(), False),
    StructField('tolls', FloatType(), False),
    StructField('total_fare', FloatType(), False)
])

# Create a DataFrame from the nycTaxiFares.gz file
taxi_fares = spark.read.csv(
    'hdfs:///learning/data/nycTaxi/nycTaxiFares.csv',
    schema=fares_schema)

# Define the schema of the nycTaxiRides dataset
rides_schema = StructType([
    StructField('ride_id', IntegerType(), False),
    StructField('is_start', StringType(), False),
    StructField('end_time', TimestampType(), False),
    StructField('start_time', TimestampType(), False),
    StructField('start_lon', FloatType(), False),
    StructField('start_lat', FloatType(), False),
    StructField('end_lon', FloatType(), False),
    StructField('end_lat', FloatType(), False),
    StructField('passenger_count', IntegerType(), False),
    StructField('taxi_id', IntegerType(), False),
    StructField('driver_id', IntegerType(), False)
])

# Create a DataFrame from the nycTaxiRides.gz file
taxi_rides = spark.read.csv(
    'hdfs:///education/ece/spark/labs/2/nycTaxiRides.gz',
    schema=rides_schema)

taxi_rides_end = taxi_rides.where('is_start = "END"').drop('taxi_id', 'driver_id', 'start_time'

complete_rides = taxi_fares.join(taxi_rides_end, 'ride_id').drop('is_start')
```

Took 0 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:01:14 AM.

SPARK JOB (http://wrk-3.au.adaltas.cloud:37393/jobs/job?id=117) FINISHED

```
%pyspark

sample = complete_rides.where('start_lon <> 0').limit(1000)

sample.show(10)
```

```
+-------+----------+----------+-------------------+------------+---+-----+----------+-----------
--------+----------+---------+----------+---------+---------------+
|ride_id|   taxi_id| driver_id|         start_time|payment_type|tip|tolls|total_fare|
end_time| start_lon|start_lat|   end_lon|  end_lat|passenger_count|
+-------+----------+----------+-------------------+------------+---+-----+----------+-----------
--------+----------+---------+----------+---------+---------------+
|    243|2013000243|2013000243|2013-01-01 00:02:00|         CRD|2.2|  0.0|      13.7|2013-01-01
00:16:00|-73.981285| 40.72491|  -73.9861|40.744987|              6|
|    392|2013000391|2013000389|2013-01-01 00:03:00|         CRD|4.8|  0.0|      29.3|2013-01-01
00:26:00| -73.92801| 40.76956| -74.00535|40.740257|              1|
|    540|2013000539|2013000537|2013-01-01 00:03:38|         CSH|0.0|  0.0|       7.0|2013-01-01
00:08:47| -73.96055|40.715336| -73.94102|40.716305|              1|
|    623|2013000622|2013000619|2013-01-01 00:04:00|         CSH|0.0|  0.0|       7.0|2013-01-01
00:10:00| -73.98825| 40.74917| -73.99686|40.757633|              1|
|    737|2013000735|2013000732|2013-01-01 00:04:36|         CRD|2.0|  0.0|      15.5|2013-01-01
00:14:33|-73.983505|  40.7302|-73.956955|40.777496|              1|
|    858|2013000853|2013000850|2013-01-01 00:05:00|         CSH|0.0|  0.0|      11.0|2013-01-01
00:14:00| -73.98802|40.738567| -74.00801|  40.7077|              1|
```

dsti-a19/ajourdan/lab5

Took 29 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:45:54 AM.

```
%pyspark

from pyspark.ml.feature import VectorAssembler

cols = ["start_lat","start_lon"]
assembler = VectorAssembler(inputCols=cols,outputCol="features")
featureDf=assembler.transform(sample)
```

FINISHED

Took 1 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:46:06 AM.

```
%pyspark
featureDf.show(10)
```

SPARK JOB (http://wrk-3.au.adaltas.cloud:37393/jobs/job?id=118) FINISHED

```
+-------+----------+----------+-------------------+------------+---+-----+----------+-----------
--------+----------+---------+----------+---------+---------------+--------------------+
|ride_id|   taxi_id| driver_id|         start_time|payment_type|tip|tolls|total_fare|
end_time| start_lon|start_lat|   end_lon|  end_lat|passenger_count|            features|
+-------+----------+----------+-------------------+------------+---+-----+----------+-----------
--------+----------+---------+----------+---------+---------------+--------------------+
|    243|2013000243|2013000243|2013-01-01 00:02:00|         CRD|2.2|  0.0|      13.7|2013-01-01
00:16:00|-73.981285| 40.72491|  -73.9861|40.744987|              6|[40.7249107360839...|
|    392|2013000391|2013000389|2013-01-01 00:03:00|         CRD|4.8|  0.0|      29.3|2013-01-01
00:26:00| -73.92801| 40.76956| -74.00535|40.740257|              1|[40.7695617675781...|
|    540|2013000539|2013000537|2013-01-01 00:03:38|         CSH|0.0|  0.0|       7.0|2013-01-01
00:08:47| -73.96055|40.715336| -73.94102|40.716305|              1|[40.7153358459472...|
|    623|2013000622|2013000619|2013-01-01 00:04:00|         CSH|0.0|  0.0|       7.0|2013-01-01
00:10:00| -73.98825| 40.74917| -73.99686|40.757633|              1|[40.7491683959960...|
|    737|2013000735|2013000732|2013-01-01 00:04:36|         CRD|2.0|  0.0|      15.5|2013-01-01
00:14:33|-73.983505|  40.7302|-73.956955|40.777496|              1|[40.7302017211914...|
|    858|2013000853|2013000850|2013-01-01 00:05:00|         CSH|0.0|  0.0|      11.0|2013-01-01
00:14:00| -73.98802|40.738567| -74.00801|  40.7077|              1|[40.7385673522949...|
```

Took 27 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:46:35 AM.

```
%pyspark
featureDf.schema
```

FINISHED

```
StructType(List(StructField(ride_id,IntegerType,true),StructField(taxi_id,IntegerType,true),StructF
ield(driver_id,IntegerType,true),StructField(start_time,TimestampType,true),StructField(payment_typ
e,StringType,true),StructField(tip,FloatType,true),StructField(tolls,FloatType,true),StructField(to
tal_fare,FloatType,true),StructField(end_time,TimestampType,true),StructField(start_lon,FloatType,t
rue),StructField(start_lat,FloatType,true),StructField(end_lon,FloatType,true),StructField(end_lat,
FloatType,true),StructField(passenger_count,IntegerType,true),StructField(features,VectorUDT,tru
e)))
```

Took 0 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:53:46 AM.

```
%pyspark

from pyspark.ml.clustering import KMeans
from pyspark.ml.evaluation import ClusteringEvaluator
```

SPARK JOBS FINISHED

```
# Trains a k-means model.
kmeans = KMeans().setK(10).setSeed(1)
model = kmeans.fit(featureDf)

# Make predictions
predictions = model.transform(featureDf)
```

dsti-a19/ajourdan/lab5

Took 11 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:46:49 AM.

SPARK JOBS FINISHED

```
%pyspark

# Evaluate clustering by computing Silhouette score
evaluator = ClusteringEvaluator()

silhouette = evaluator.evaluate(predictions)
print("Silhouette with squared euclidean distance = " + str(silhouette))

# Shows the result.
centers = model.clusterCenters()
print("Cluster Centers: ")
for center in centers:
    print(center)
```

```
Silhouette with squared euclidean distance = 0.484830036365
Cluster Centers: 
[ 40.74499732 -73.98266447]
[ 40.7389471  -74.00054232]
[ 40.67258606 -73.7977478 ]
[ 40.776328   -73.88172358]
[ 40.7035719 -73.9547761]
[ 40.72034029 -73.99389566]
[ 40.82387543 -74.12619781]
[ 40.79420509 -73.96341232]
[ 40.76563301 -73.98424389]
[ 40.76849702 -73.95834917]
```

Took 30 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:47:31 AM.

SPARK JOB (http://wrk-3.au.adaltas.cloud:37393/jobs/job?id=153) FINISHED

```
%pyspark
predictions.show(10)
```

```
+-------+----------+----------+-------------------+------------+---+-----+----------+-----------
--------+----------+---------+----------+---------+---------------+-------------------+--------
--+
|ride_id|   taxi_id| driver_id|         start_time|payment_type|tip|tolls|total_fare|
end_time| start_lon|start_lat|   end_lon|  end_lat|passenger_count|           features|predicti
on|
+-------+----------+----------+-------------------+------------+---+-----+----------+-----------
--------+----------+---------+----------+---------+---------------+-------------------+--------
--+
|    148|2013000148|2013000148|2013-01-01 00:01:00|         CRD|5.6|  0.0|      34.1|2013-01-01
00:31:00| -73.95757|40.722225|  -73.9823|40.768288|              6|[40.7222251892089...|
4|
|    463|2013000462|2013000460|2013-01-01 00:03:00|         CSH|0.0|  0.0|      12.0|2013-01-01
00:17:00| -73.99193|40.721977| -73.97819|40.745796|              1|[40.7219772338867...|
5|
|    471|2013000470|2013000468|2013-01-01 00:03:00|         CSH|0.0|  0.0|      10.5|2013-01-01
00:10:00| -73.99046|40.731068| -73.96442| 40.75617|              2|[40.7310676574707...|
5|
```

Took 31 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:48:21 AM.

```
%pyspark
pred_groups = predictions \
                .groupBy("prediction") \
                .agg(
                    F.count("ride_id").alias("nb_ride")
                    ) \
                .orderBy("prediction")


pred_groups.show(10)
```

SPARK JOBS FINISHED

```
+----------+-------+
|prediction|nb_ride|
+----------+-------+
|         0|    191|
|         1|    148|
|         2|      5|
|         3|     11|
|         4|     47|
|         5|    159|
|         6|      1|
|         7|    100|
|         8|    152|
|         9|    186|
+----------+-------+
```

Took 10 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:48:34 AM.

```
%pyspark
```

FINISHED

Took 24 sec. Last updated by a.jourdan-dsti at March 19 2020, 11:41:52 AM. (outdated)

```
%pyspark
```

READY