Spark Project – 1 (Taxi App Company)

Loading the data

```
scala> val taxiDF = (spark
     | .read
     | .format("csv")
     | .option("inferSchema","true")
     | .option("header","true")
     | .load("/home/bitnami/sparkdata/yellow.csv"))
taxiDF: org.apache.spark.sql.DataFrame = [VendorID: int, tpep_pickup_datetime: string ... 17 more fields
]
```

Viewing first few records

```
scala> taxiDF.show(5)
+--------+--------------------+--------------------+---------------+-------------+-----------------+---
-------------+----------+-----------------+-----------------+---------------+------------+-----------+
-----+-------+----------+------------+------------+---------+
|VendorID|tpep_pickup_datetime|tpep_dropoff_datetime|passenger_count|trip_distance| pickup_longitude| pi
ckup_latitude|RateCodeID|store_and_fwd_flag|dropoff_longitude|dropoff_latitude|payment_type|fare_amount|
extra|mta_tax|tip_amount|tolls_amount|total_amount|trip_time|
+--------+--------------------+--------------------+---------------+-------------+-----------------+---
-------------+----------+-----------------+-----------------+---------------+------------+-----------+
-----+-------+----------+------------+------------+---------+
|       2| 2015-01-08 22:44:09|  2015-01-08 22:50:56|              1|         1.55|-73.9876861572266| 40
.724250793457|         1|                N| -73.973762512207|40.7433776855469|           2|        7.5|
  0.5|    0.5|       0.0|         0.0|         8.8|     5000|
|       1| 2015-01-08 22:44:09|  2015-01-08 22:51:17|              3|          1.2| -73.991569519043|40.
7269325256348|         1|                N|-74.0041046142578|40.7210807800293|           2|        7.0|
  0.5|    0.5|       0.0|         0.0|         8.3|  5344860|
|       1| 2015-01-08 22:44:10|  2015-01-08 22:55:27|              1|          2.4|-73.9819183349609|40.
7834434509277|         1|                N|-73.9523544311524|40.7981986999512|           2|       10.5|
  0.5|    0.5|       0.0|         0.0|        11.8|  3345464|
|       1| 2015-01-08 22:44:10|  2015-01-08 22:58:09|              1|          7.3|-73.9731216430664|40.
7435531616211|         1|                N|-73.9195709228516|40.8320007324219|           2|       21.5|
  0.5|    0.5|       0.0|         0.0|        22.8|   893933|
|       1| 2015-01-08 22:44:12|  2015-01-08 22:46:16|              1|          0.4|-73.9829483032227|40.
7662086486816|         1|                N|-73.9843902587891|40.7640533447266|           2|        3.5|
  0.5|    0.5|       0.0|         0.0|         4.8|    36864|
+--------+--------------------+--------------------+---------------+-------------+-----------------+---
-------------+----------+-----------------+-----------------+---------------+------------+-----------+
-----+-------+----------+------------+------------+---------+
only showing top 5 rows
```

Creating a View

```
scala> taxiDF.createOrReplaceTempView("tblTaxiData")
```

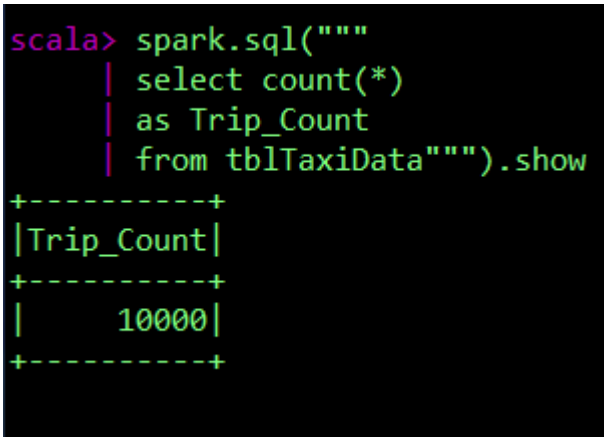
Viewing the data after creating a view

```
scala> spark.sql("select * from tblTaxiData").show(5)
+--------+--------------------+---------------------+---------------+-------------+-----------------+---
------------+----------+-----------------+------------+-----------------+-------------+-----------+
-----+-------+----------+------------+------------+---------+
|VendorID|tpep_pickup_datetime|tpep_dropoff_datetime|passenger_count|trip_distance| pickup_longitude| pi
ckup_latitude|RateCodeID|store_and_fwd_flag|dropoff_longitude|dropoff_latitude|payment_type|fare_amount|
extra|mta_tax|tip_amount|tolls_amount|total_amount|trip_time|
+--------+--------------------+---------------------+---------------+-------------+-----------------+---
------------+----------+-----------------+------------+-----------------+-------------+-----------+
-----+-------+----------+------------+------------+---------+
|       2| 2015-01-08 22:44:09|  2015-01-08 22:50:56|              1|         1.55|-73.9876861572266| 40
.724250793457|         1|                N| -73.973762512207|40.7433776855469|           2|        7.5|
  0.5|    0.5|       0.0|         0.0|         8.8|     5000|
|       1| 2015-01-08 22:44:09|  2015-01-08 22:51:17|              3|          1.2| -73.991569519043|40.
7269325256348|         1|                N|-74.0041046142578|40.7210807800293|           2|        7.0|
  0.5|    0.5|       0.0|         0.0|         8.3|  5344860|
|       1| 2015-01-08 22:44:10|  2015-01-08 22:55:27|              1|          2.4|-73.9819183349609|40.
7834434509277|         1|                N|-73.9523544311524|40.7981986999512|           2|       10.5|
  0.5|    0.5|       0.0|         0.0|        11.8|  3345464|
|       1| 2015-01-08 22:44:10|  2015-01-08 22:58:09|              1|          7.3|-73.9731216430664|40.
7435531616211|         1|                N|-73.9195709228516|40.8320007324219|           2|       21.5|
  0.5|    0.5|       0.0|         0.0|        22.8|   893933|
|       1| 2015-01-08 22:44:12|  2015-01-08 22:46:16|              1|          0.4|-73.9829483032227|40.
7662086486816|         1|                N|-73.9843902587891|40.7640533447266|           2|        3.5|
  0.5|    0.5|       0.0|         0.0|         4.8|    36864|
+--------+--------------------+---------------------+---------------+-------------+-----------------+---
------------+----------+-----------------+------------+-----------------+-------------+-----------+
-----+-------+----------+------------+------------+---------+
only showing top 5 rows
```

Use Spark shell to perform the following tasks:

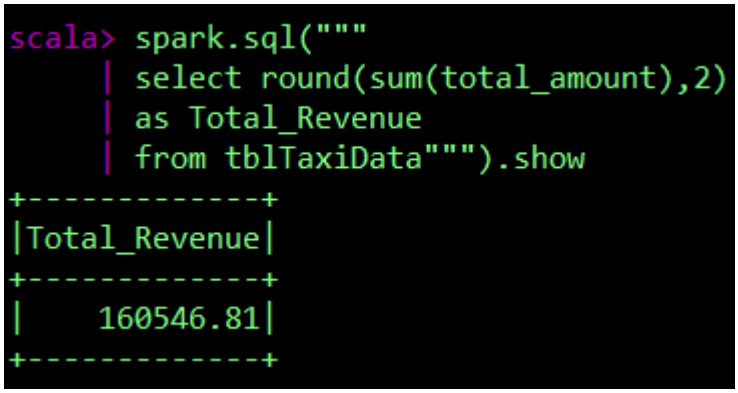
1. What is the total number of trips (equal to the number of rows)?

Solution:



2. What is the total revenue generated by all the trips? The fare is stored in the column, total_amount.

Solution:



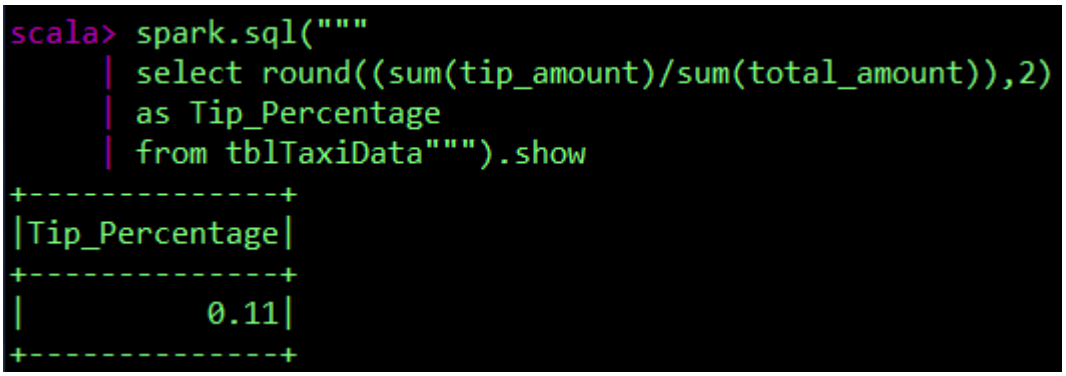
3. What fraction of the total is paid for tolls? The toll is stored in tolls_amount.

Solution:

```
scala> spark.sql("""
     | select round((sum(tolls_amount)/sum(total_amount)),2)
     | as Toll_Percentage
     | from tblTaxiData""").show
+---------------+
|Toll_Percentage|
+---------------+
|           0.02|
+---------------+
```

4. What fraction of it is given as driver tips? The tip is stored in tip_amount.

Solution:



5. What is the average trip amount?

Solution:

```
scala> spark.sql("""
     | select round(avg(total_amount),2)
     | as Avg_Trip_Amount
     | from tblTaxiData""").show
+---------------+
|Avg_Trip_Amount|
+---------------+
|          16.05|
+---------------+
```

6. What is the average distance of the trips? Distance is stored in the column, trip_distance.

Solution:

```
scala> spark.sql("""
     | select round(avg(trip_distance),2)
     | as Avg_Trip_Distance
     | from tblTaxiData""").show
+-----------------+
|Avg_Trip_Distance|
+-----------------+
|             3.25|
+-----------------+
```

7. How many different payment types are used? AND
8. For each payment type, display the following details: a. Average fare generated b. Average tip c. Average tax – tax is stored in the column, mta_tax.

Solution:

```
scala> spark.sql("""
     | select payment_type,
     | round(avg(fare_amount),2) as Average_Fare_Amount,
     | round(avg(tip_amount),2) as Average_Tip_Amount,
     | round(avg(mta_tax),2) as Average_Tax_Amount
     | from tblTaxiData
     | group by payment_type""").show
+------------+-------------------+------------------+------------------+
|payment_type|Average_Fare_Amount|Average_Tip_Amount|Average_Tax_Amount|
+------------+-------------------+------------------+------------------+
|           1|              13.56|               2.7|               0.5|
|           3|              13.21|               0.0|              0.42|
|           4|              12.22|               0.0|               0.5|
|           2|              11.39|               0.0|               0.5|
+------------+-------------------+------------------+------------------+
```

9. On average, which hour of the day generates the highest revenue?

Solution:

```
scala> spark.sql("""
     | select h24 as hour,
     | round(avg(total_amount),2) as Average_Revenue
     | from (select hour(tpep_pickup_datetime) as h24,
     | total_amount
     | from tblTaxiData)
     | ff
     | group by h24
     | order by Average_Revenue desc""").show
+----+---------------+
|hour|Average_Revenue|
+----+---------------+
|  22|          16.24|
|  23|          16.11|
|   0|          15.32|
+----+---------------+
```