Soudabeh Rafieisakhaei

1. Create a subdirectory SPRK/Ex2 in HDFS and upload the d311.csv file to that subdirectory.

   A: Creating a subdirectory SPRK/Ex2 in HDFS

```
[hadoop@ip-172-31-16-163 ~]$ hadoop fs -mkdir SPRK
[hadoop@ip-172-31-16-163 ~]$ hadoop fs -mkdir SPRK/EX2
```

   B: Uploading the d311.csv file to that subdirectory.

```
[hadoop@ip-172-31-16-163 ~]$ hadoop fs -copyFromLocal d311.csv SPRK/EX2
[hadoop@ip-172-31-16-163 ~]$
```

2. Start the Spark shell and read the d311.csv file. View the schema and note that the column names match the record field names in the CSV file. Provide a screenshot of the schema.

```
[hadoop@ip-172-31-16-163 ~]$ spark-shell
```

```
scala> val EX2DF=spark.read.format("csv").option("header","true").load("SPRK/EX2/d311.csv")
EX2DF: org.apache.spark.sql.DataFrame = [Unique Key: string, Created Date: string ... 36 more fields]
```

```
scala> EX2DF.printSchema()
root
 |-- Unique Key: string (nullable = true)
 |-- Created Date: string (nullable = true)
 |-- Closed Date: string (nullable = true)
 |-- Agency: string (nullable = true)
 |-- Agency Name: string (nullable = true)
 |-- Complaint Type: string (nullable = true)
 |-- Descriptor: string (nullable = true)
 |-- Location Type: string (nullable = true)
 |-- Incident Zip: string (nullable = true)
 |-- Incident Address: string (nullable = true)
 |-- Street Name: string (nullable = true)
 |-- Cross Street 1: string (nullable = true)
 |-- Cross Street 2: string (nullable = true)
 |-- Intersection Street 1: string (nullable = true)
 |-- Intersection Street 2: string (nullable = true)
 |-- Address Type: string (nullable = true)
 |-- City: string (nullable = true)
 |-- Landmark: string (nullable = true)
 |-- Facility Type: string (nullable = true)
 |-- Status: string (nullable = true)
 |-- Due Date: string (nullable = true)
 |-- Resolution Action Updated Date: string (nullable = true)
 |-- Community Board: string (nullable = true)
 |-- Borough: string (nullable = true)
 |-- X Coordinate (State Plane): string (nullable = true)
 |-- Y Coordinate (State Plane): string (nullable = true)
 |-- Park Facility Name: string (nullable = true)
 |-- Park Borough: string (nullable = true)
 |-- Vehicle Type: string (nullable = true)
 |-- Taxi Company Borough: string (nullable = true)
 |-- Taxi Pick Up Location: string (nullable = true)
 |-- Bridge Highway Name: string (nullable = true)
 |-- Bridge Highway Direction: string (nullable = true)
 |-- Road Ramp: string (nullable = true)
 |-- Bridge Highway Segment: string (nullable = true)
 |-- Latitude: string (nullable = true)
 |-- Longitude: string (nullable = true)
 |-- Location: string (nullable = true)


scala>
```

3. Display the data in the DataFrame using the show function. How many records are displayed? Display the first five records of the data frame. Provide a screenshot of the result.

A: `scala> EX2DF.show()` Only 20 rows of records displayed.

B: Using the show function to show the first 5 records:

```
scala> EX2DF.show(5)
+----------+------------+------------+------+-------------------+-------------------+-------------------+-------------------+-----------+------------
------+----------------+--------------+------------------+-------------------+-------------------+------------+---------------+-------+------------
+------+------------+-------------------------+-------------------+--------+-----------+-------------------+-------------------+-----------------+--
----------+------------+-------------------+-------------------+-------------------+--------------------+--------+-------------------+-----------+--
----------+-------------------+
|Unique Key| Created Date|   Closed Date|Agency|         Agency Name|      Complaint Type|          Descriptor|       Location Type|Incident Zip|    Incident A
ddress|      Street Name|   Cross Street 1|      Cross Street 2|Intersection Street 1|Intersection Street 2|Address Type|          City|Landmark|Facility Type
|Status|      Due Date|Resolution Action Updated Date|      Community Board|  Borough|X Coordinate (State Plane)|Y Coordinate (State Plane)|Park Facility Name|Pa
rk Borough|Vehicle Type|Taxi Company Borough|Taxi Pick Up Location|Bridge Highway Name|Bridge Highway Direction|Road Ramp|Bridge Highway Segment|    Latitude|
 Longitude|             Location|
+----------+------------+------------+------+-------------------+-------------------+-------------------+-------------------+-----------+------------
------+----------------+--------------+------------------+-------------------+-------------------+------------+---------------+-------+------------
+------+------------+-------------------------+-------------------+--------+-----------+-------------------+-------------------+-----------------+--
----------+------------+-------------------+-------------------+-------------------+--------------------+--------+-------------------+-----------+--
----------+-------------------+
|  28163399|6/1/2014 0:00| 6/1/2014 0:15|   DOT|Department of Tra...|Traffic Signal Co...| LED Pedestrian Unit|                NULL|        NULL|
  NULL|             NULL|             NULL|                NULL|                18 AVE|                4 ST E|INTERSECTION|          NULL|    NULL|          N/A
|Closed|         NULL|                6/1/2014 0:15|Unspecified BROOKLYN| BROOKLYN|                NULL|                        NULL|       Unspecified|
  BROOKLYN|        NULL|                NULL|                NULL|                NULL|                NULL|     NULL|                NULL|        NULL|
      NULL|                 NULL|
|  28157590|6/1/2014 0:00|6/6/2014 16:03| DOHMH|Department of Hea...|              Rodent|      Mouse Sighting| 1-2 Family Dwelling|       11372|70-04 ROOSEVEL
T A...| ROOSEVELT AVENUE|        70 STREET|BQE WESTBOUND ENT...|                NULL|                NULL|     ADDRESS|Jackson Heights|    NULL|          N/A
|Closed|7/1/2014 0:45|               6/6/2014 16:03|           02 QUEENS|   QUEENS|             1013248|                      211238|       Unspecified|
    QUEENS|        NULL|                NULL|                NULL|                NULL|                NULL|     NULL|                NULL|40.74642855| -
73.8953472|(40.7464285492033...|
|  28157974|6/1/2014 0:00|6/10/2014 0:00|   HPD|Department of Hou...|UNSANITARY CONDITION|GARBAGE/RECYCLING...|RESIDENTIAL BUILDING|       10024|  336 WEST 77
STREET|   WEST 77 STREET|  WEST END AVENUE|     RIVERSIDE DRIVE|                NULL|                NULL|     ADDRESS|       NEW YORK|    NULL|          N/A
|Closed|         NULL|               6/10/2014 0:00|        07 MANHATTAN|MANHATTAN|              988998|                      224663|       Unspecified|
 MANHATTAN|        NULL|                NULL|                NULL|                NULL|                NULL|     NULL|                NULL|40.78332298|-7
3.98285514|(40.7833229754180...|
|  28158733|6/1/2014 0:00| 6/9/2014 0:00|   HPD|Department of Hou...|UNSANITARY CONDITION|               PESTS|RESIDENTIAL BUILDING|       11355|   140-37 ASH
AVENUE|       ASH AVENUE|KISSENA BOULEVARD|        BOWNE STREET|                NULL|                NULL|     ADDRESS|       Flushing|    NULL|          N/A
|Closed|         NULL|                6/9/2014 0:00|           07 QUEENS|   QUEENS|             1033351|                      214635|       Unspecified|
    QUEENS|        NULL|                NULL|                NULL|                NULL|                NULL|     NULL|                NULL|40.75566367| -
73.8227715|(40.7556636656514...|
|  28159418|6/1/2014 0:00|6/20/2014 0:00|   HPD|Department of Hou...|UNSANITARY CONDITION|                MOLD|RESIDENTIAL BUILDING|       10458|2604 BAINBRIDG
E A...|BAINBRIDGE AVENUE|  EAST 193 STREET|     EAST 194 STREET|                NULL|                NULL|     ADDRESS|          BRONX|    NULL|          N/A
|Closed|         NULL|               6/20/2014 0:00|            07 BRONX|    BRONX|             1014036|                      254171|       Unspecified|
     BRONX|        NULL|                NULL|                NULL|                NULL|                NULL|     NULL|                NULL|40.86426508|-7
3.89231286|(40.8642650819864...|
+----------+------------+------------+------+-------------------+-------------------+-------------------+-------------------+-----------+------------
------+----------------+--------------+------------------+-------------------+-------------------+------------+---------------+-------+------------
+------+------------+-------------------------+-------------------+--------+-----------+-------------------+-------------------+-----------------+--
----------+------------+-------------------+-------------------+-------------------+--------------------+--------+-------------------+-----------+--
```

4. Use the count action to return the number of items in the DataFrame. Provide a screenshot of the result.

```
scala> val count=EX2DF.count()
count: Long = 518909
```

5. Use a select transformation to return a DataFrame with only the Created Date, Agency, Complaint Type, and City. The select transformation should return all columns with an alias instead of the real name. Display the schema of the new DataFrame. Provide a screenshot of the result.

A:

```
scala> val selectedDF=EX2DF.select($"Created Date".alias("Date_Created"), $"Agency".alias("Agency"),$"Complaint Type".alias("Complaint_Type"),$"City".alias("City"))
selectedDF: org.apache.spark.sql.DataFrame = [Date_Created: string, Agency: string ... 2 more fields]
```

B:

```
scala> selectedDF.printSchema()
root
 |-- Date_Created: string (nullable = true)
 |-- Agency: string (nullable = true)
 |-- Complaint_Type: string (nullable = true)
 |-- City: string (nullable = true)
```

6. Write a query (a series of one or more transformations followed by an action) that displays the first 20 lines of Agency, City, Complaint Type, where City is not null. Provide a screenshot of the result.

without assigning the number of rows in the show function

```
scala> selectedDF.select("Agency","City", "Complaint_Type").where($"City".isNotNull).show(20)
+------+---------------+--------------------+
|Agency|           City|      Complaint_Type|
+------+---------------+--------------------+
| DOHMH|Jackson Heights|              Rodent|
|   HPD|       NEW YORK|UNSANITARY CONDITION|
|   HPD|       Flushing|UNSANITARY CONDITION|
|   HPD|          BRONX|UNSANITARY CONDITION|
|   HPD|       NEW YORK|UNSANITARY CONDITION|
|   HPD|       Flushing|          WATER LEAK|
|   HPD|       Flushing|          WATER LEAK|
|   HPD|          BRONX|UNSANITARY CONDITION|
|   HPD|       Flushing|          WATER LEAK|
| DOHMH|       BROOKLYN|              Rodent|
| DOHMH|       Flushing|              Rodent|
|   HPD|       BROOKLYN|      HEAT/HOT WATER|
| DOHMH|       BROOKLYN|              Rodent|
| DOHMH|          BRONX|              Rodent|
| DOHMH|          BRONX|              Rodent|
| DOHMH|          BRONX|              Rodent|
| DOHMH|          BRONX|              Rodent|
| DOHMH|       BROOKLYN|              Rodent|
| DOHMH|       BROOKLYN|              Rodent|
| DOHMH|       NEW YORK|              Rodent|
+------+---------------+--------------------+
only showing top 20 rows
```

Or

```
scala> val Q6=selectedDF.select("Agency", "City", "Complaint_Type").where($"City".isNotNull)
Q6: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Agency: string, City: string ... 1 more field]
```

```
scala> Q6.show(20)
+------+--------------+--------------------+
|Agency|          City|      Complaint_Type|
+------+--------------+--------------------+
| DOHMH|Jackson Heights|              Rodent|
|   HPD|      NEW YORK|UNSANITARY CONDITION|
|   HPD|      Flushing|UNSANITARY CONDITION|
|   HPD|         BRONX|UNSANITARY CONDITION|
|   HPD|      NEW YORK|UNSANITARY CONDITION|
|   HPD|      Flushing|          WATER LEAK|
|   HPD|      Flushing|          WATER LEAK|
|   HPD|         BRONX|UNSANITARY CONDITION|
|   HPD|      Flushing|          WATER LEAK|
| DOHMH|      BROOKLYN|              Rodent|
| DOHMH|      Flushing|              Rodent|
|   HPD|      BROOKLYN|      HEAT/HOT WATER|
| DOHMH|      BROOKLYN|              Rodent|
| DOHMH|         BRONX|              Rodent|
| DOHMH|         BRONX|              Rodent|
| DOHMH|         BRONX|              Rodent|
| DOHMH|         BRONX|              Rodent|
| DOHMH|      BROOKLYN|              Rodent|
| DOHMH|      BROOKLYN|              Rodent|
| DOHMH|      NEW YORK|              Rodent|
+------+--------------+--------------------+
only showing top 20 rows
```