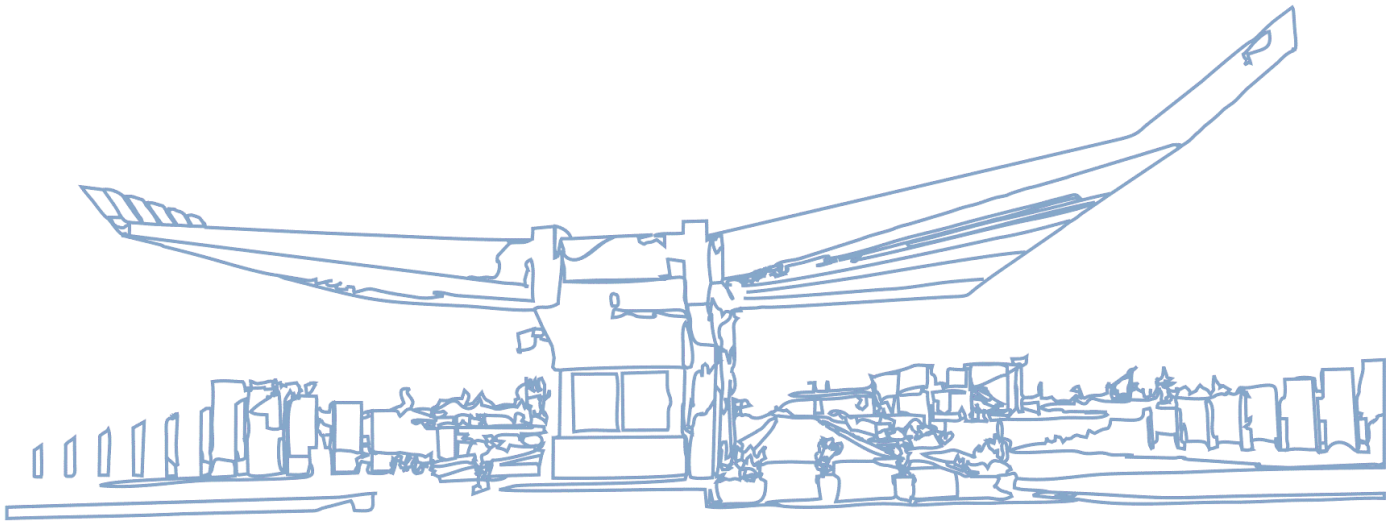
EPOKA UNIVERSITY | Computer Engineering Department

CEN 571 – Data Mining

Assignment 6: DataFrames



PREPARED:
Baftjar TABAKU

24.05.2020
Epoka University
Tirana, ALBANIA

ACCEPTED:
Prof.Dr. Arben Asllani

Tasks to completeUse d311.csv (source: https://data.cityofnewyork.us/Social-Services/311-csv/i4zx-95p9) file as a data source and perform the following steps:

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

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'. Provide a screenshot of the schema.

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

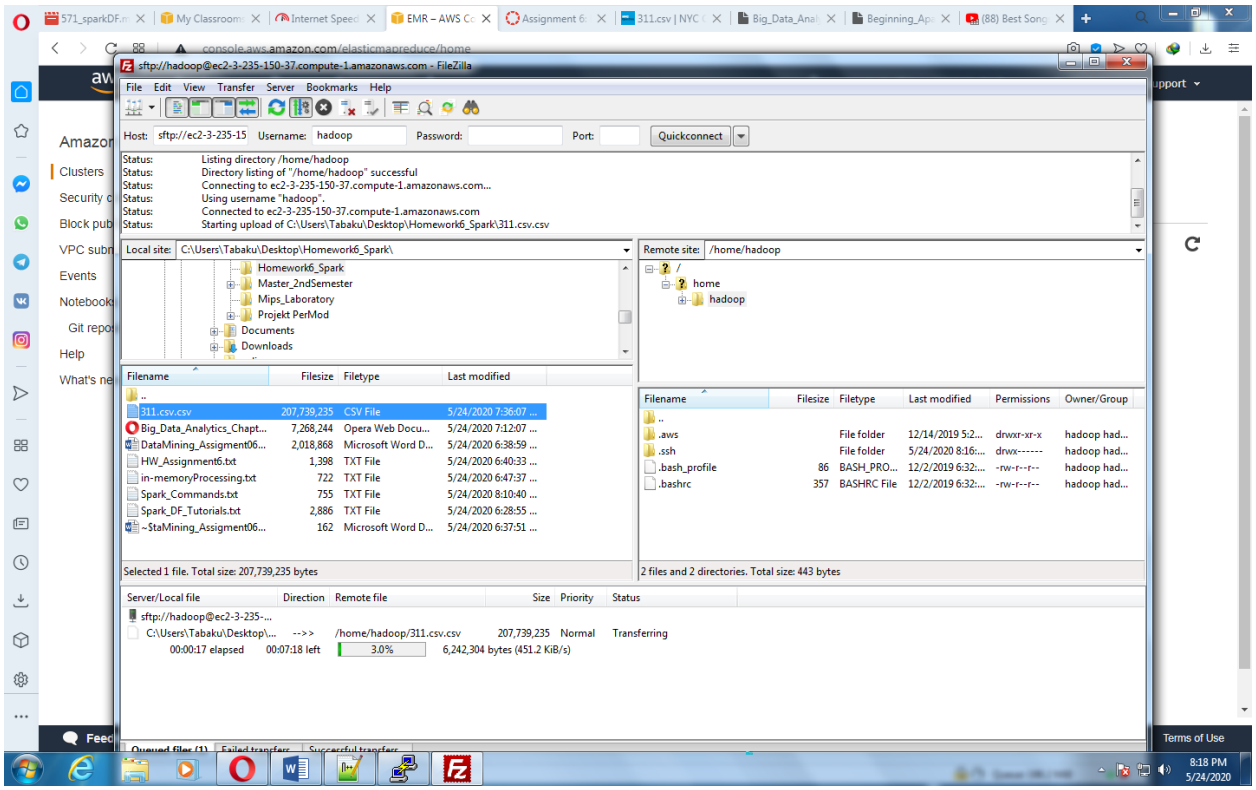
4. Use the count action to return the number of items in the DataFrame. Provide a screenshot of the result.

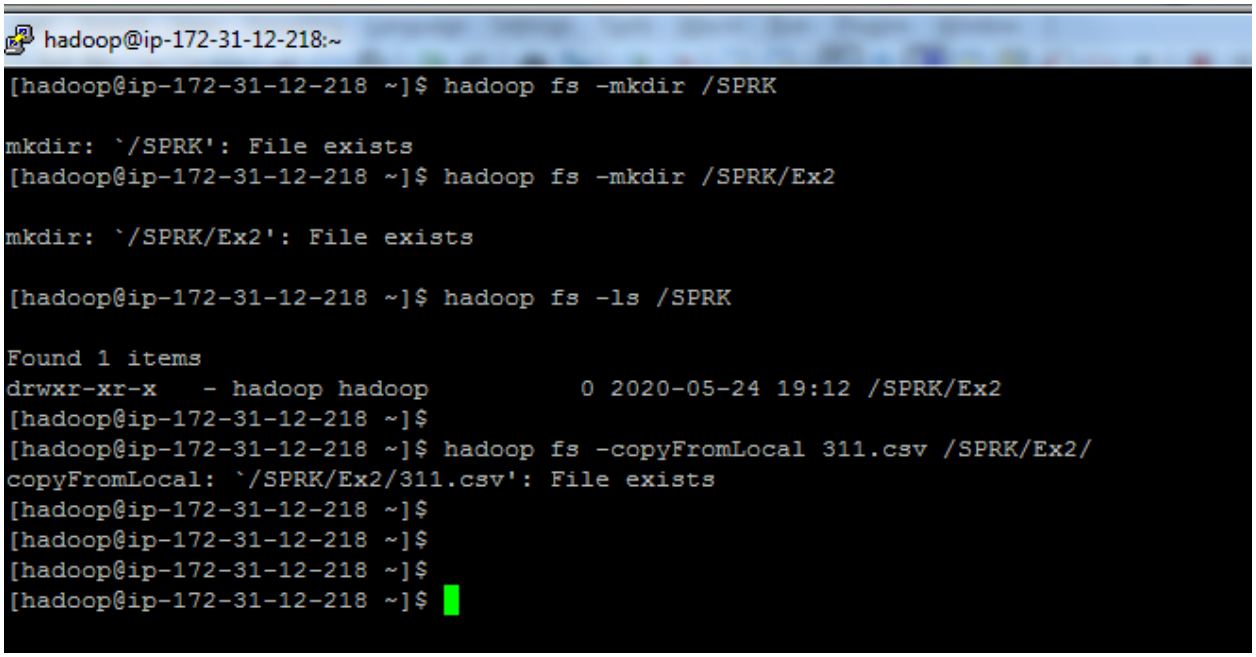
5. Using a select transformation to return a DataFrame with only the ‘Created Date’, ‘Agency’, ‘Complaint Type’ and ‘City’. The select transformation should return all columns with an alias instead of the real name. Display the schema of the new DataFrame. Provide a screenshot of the result.

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

7. Perform the same query as in #6 above, but this time execute a single command to show the same results. Provide a screenshot of the result.

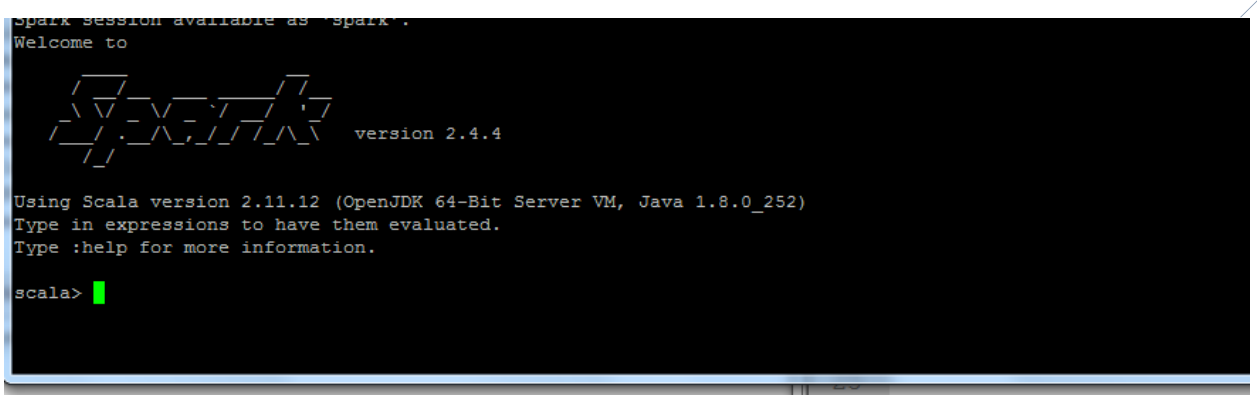
1. After successfully uploading the data into the Hadoop cluster we proceed on further analyze.

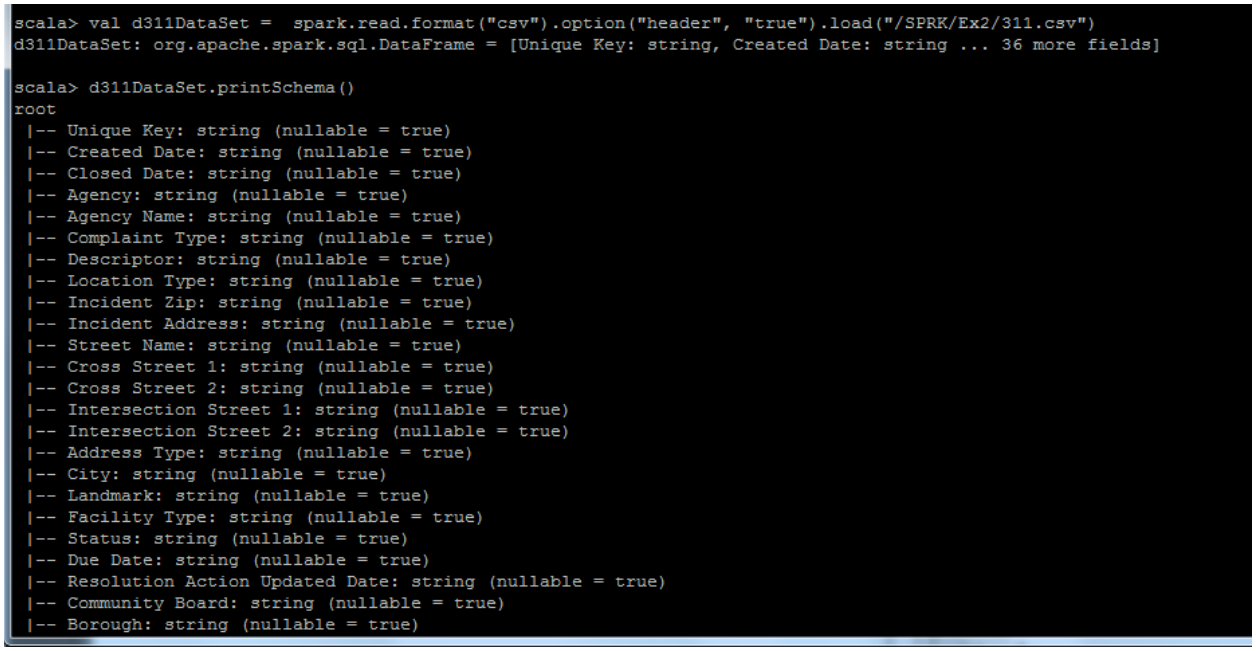




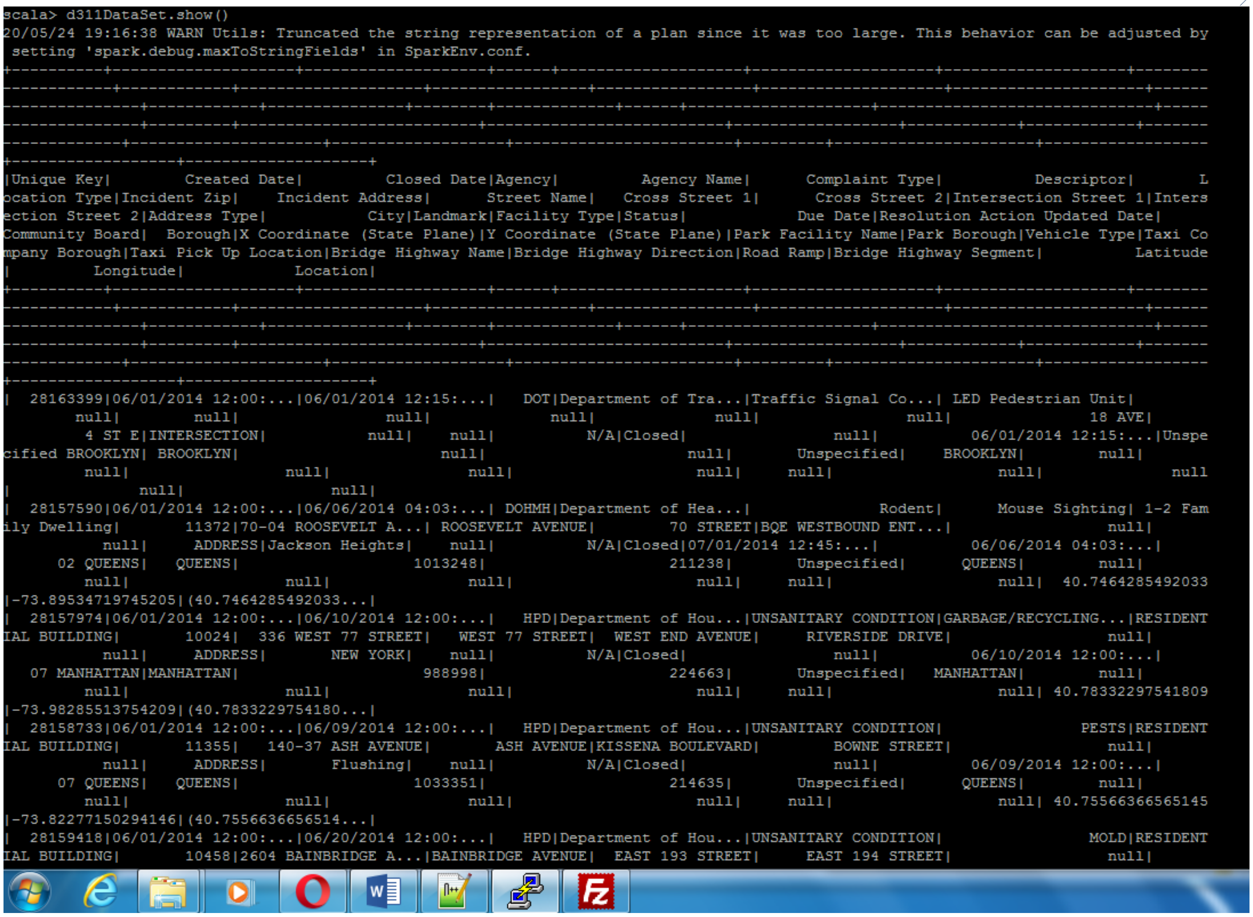
.

2. Staring the spark shell and proceeding further as shown below

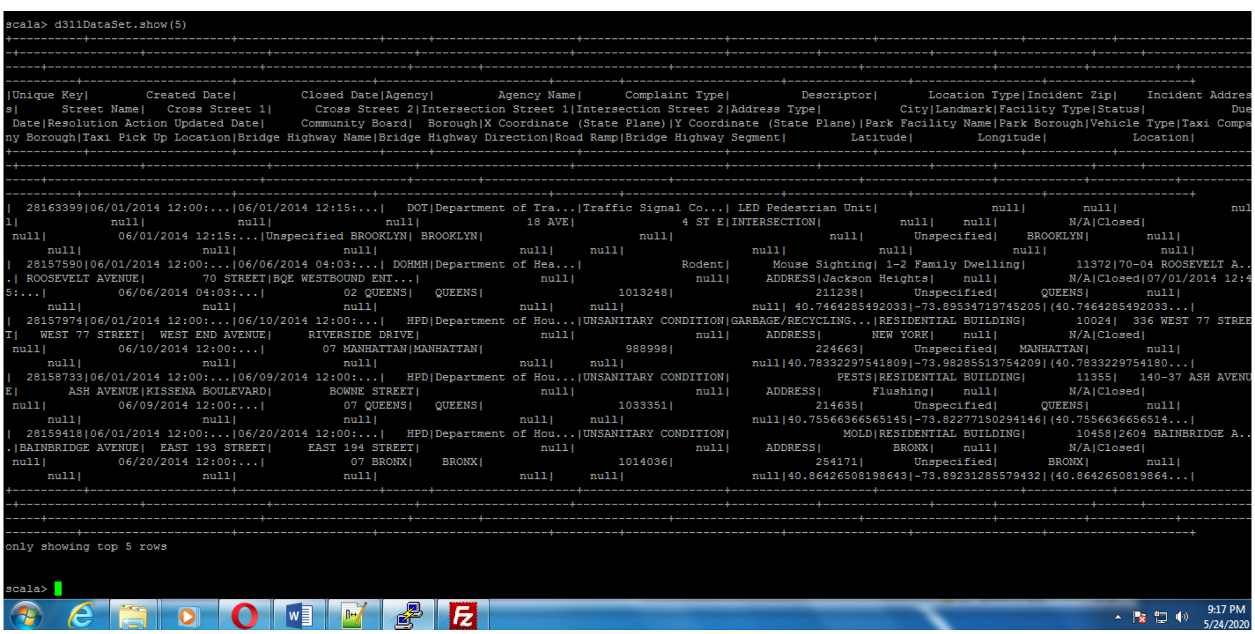




3. Showing the data,



Showing the first five records



Note: it shows a mess because the screen is small.

4. Counting data

```
scala> val numCount = d311DataSet.count()
numCount: Long = 518909
```

5. Using 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.

```
scala> val costumQuerySelect = d311DataSet.select($"Created Date".alias("A1"), $"Agency".alias("A2"), $"Complaint Type".alias("A3"), $"City".alias("A4"))
costumQuerySelect: org.apache.spark.sql.DataFrame = [A1: string, A2: string ... 2 more fields]

scala> costumQuerySelect.printSchema();
root
 |-- A1: string (nullable = true)
 |-- A2: string (nullable = true)
 |-- A3: string (nullable = true)
 |-- A4: string (nullable = true)


scala> costumQuerySelect.show()
+--------------------+-----+--------------------+---------------+
|                  A1|   A2|                  A3|             A4|
+--------------------+-----+--------------------+---------------+
|06/01/2014 12:00:...|  DOT|Traffic Signal Co...|           null|
|06/01/2014 12:00:...|DOHMH|              Rodent|Jackson Heights|
|06/01/2014 12:00:...|  HPD|UNSANITARY CONDITION|       NEW YORK|
|06/01/2014 12:00:...|  HPD|UNSANITARY CONDITION|       Flushing|
|06/01/2014 12:00:...|  HPD|UNSANITARY CONDITION|          BRONX|
|06/01/2014 12:00:...|  HPD|UNSANITARY CONDITION|       NEW YORK|
|06/01/2014 12:00:...|  HPD|          WATER LEAK|       Flushing|
|06/01/2014 12:00:...|  HPD|          WATER LEAK|       Flushing|
|06/01/2014 12:00:...|  HPD|UNSANITARY CONDITION|          BRONX|
|06/01/2014 12:00:...|  HPD|          WATER LEAK|       Flushing|
|06/01/2014 12:00:...|DOHMH|              Rodent|       BROOKLYN|
|06/01/2014 12:00:...|DOHMH|              Rodent|       Flushing|
|06/01/2014 12:00:...|  HPD|      HEAT/HOT WATER|       BROOKLYN|
|06/01/2014 12:00:...|DOHMH|              Rodent|       BROOKLYN|
|06/01/2014 12:00:...|DOHMH|              Rodent|          BRONX|
|06/01/2014 12:00:...|DOHMH|              Rodent|          BRONX|
|06/01/2014 12:00:...|DOHMH|              Rodent|          BRONX|
|06/01/2014 12:00:...|DOHMH|              Rodent|          BRONX|
|06/01/2014 12:00:...|DOHMH|              Rodent|       BROOKLYN|
|06/01/2014 12:00:...|DOHMH|              Rodent|       BROOKLYN|
+--------------------+-----+--------------------+---------------+
only showing top 20 rows


scala>
```

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

```
scala> val costumQuerySelect6 = d311DataSet.select("Agency", "City", "Complaint Type").where("City is not Null")
costumQuerySelect6: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Agency: string, City: string ... 1 more field]

scala> costumQuerySelect6.show()
+------+--------------+--------------------+
|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


scala> costumQuerySelect6.printSchema()
root
 |-- Agency: string (nullable = true)
 |-- City: string (nullable = true)
 |-- Complaint Type: string (nullable = true)


scala>
```

By default it displays 20 values so no need to specify .show (20).

7. Executing the same command as exercise #6 but this time will a single command to show the same results.

```
scala> d311DataSet.select("Agency", "City", "Complaint Type").where("City is not Null").show()
+------+--------------+--------------------+
|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


scala>
```