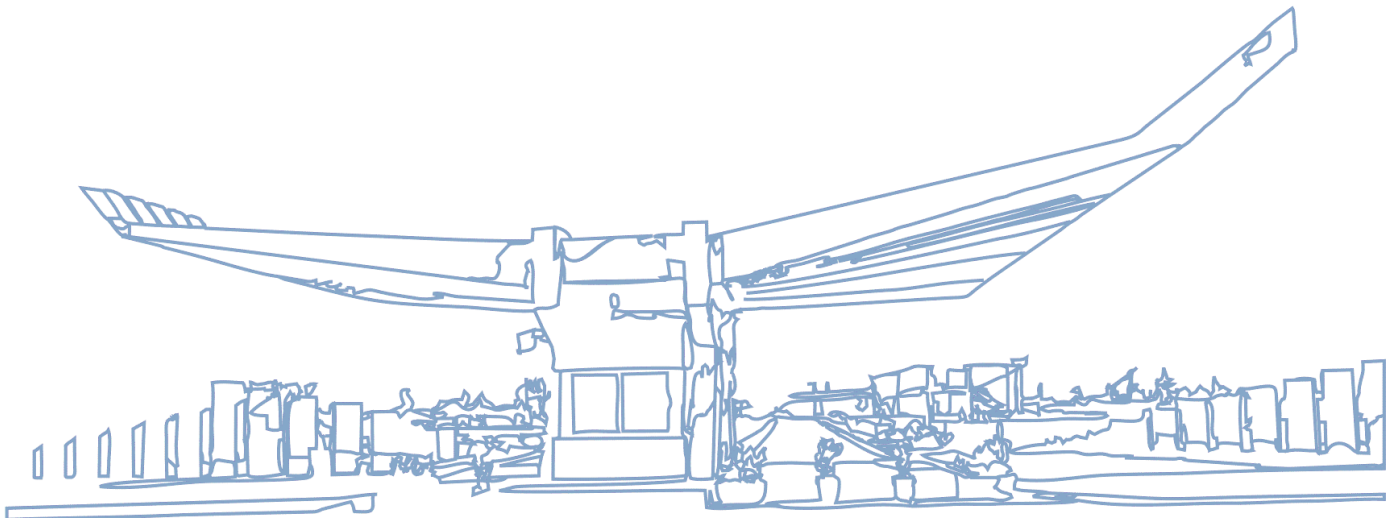


CEN 571 – Data Mining

Assignment 6: DataFrames



PREPARED:

Baftjar TABAKU

24.05.2020

Epoka University
Tirana, ALBANIA

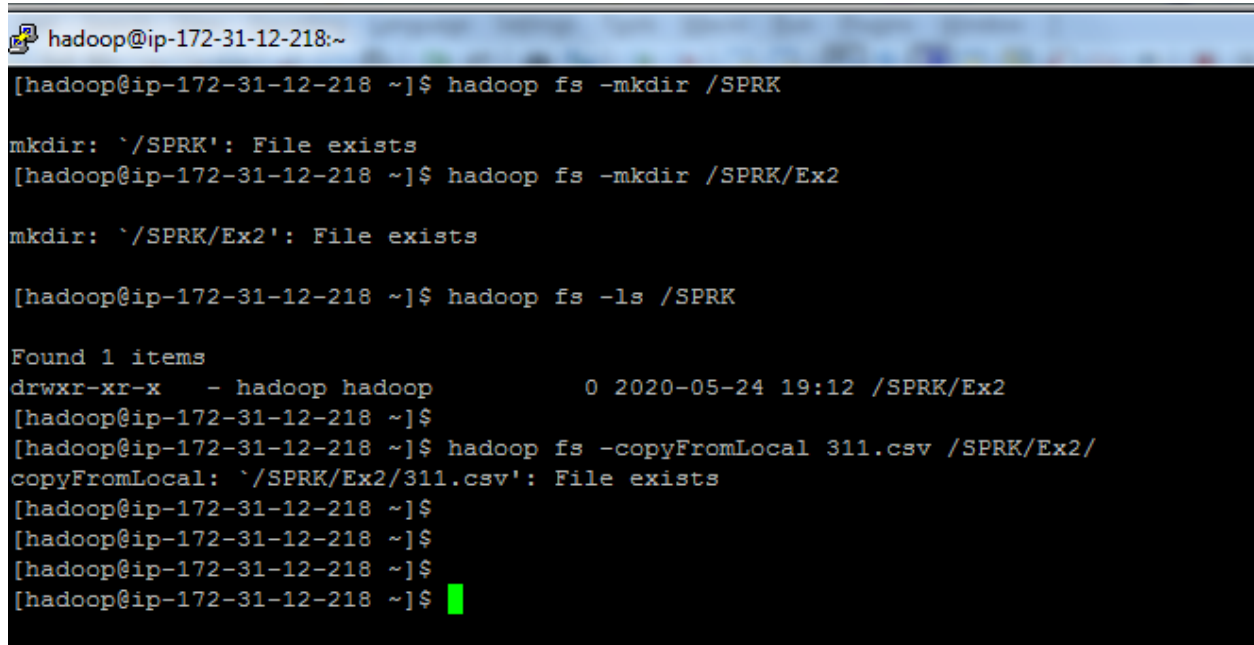
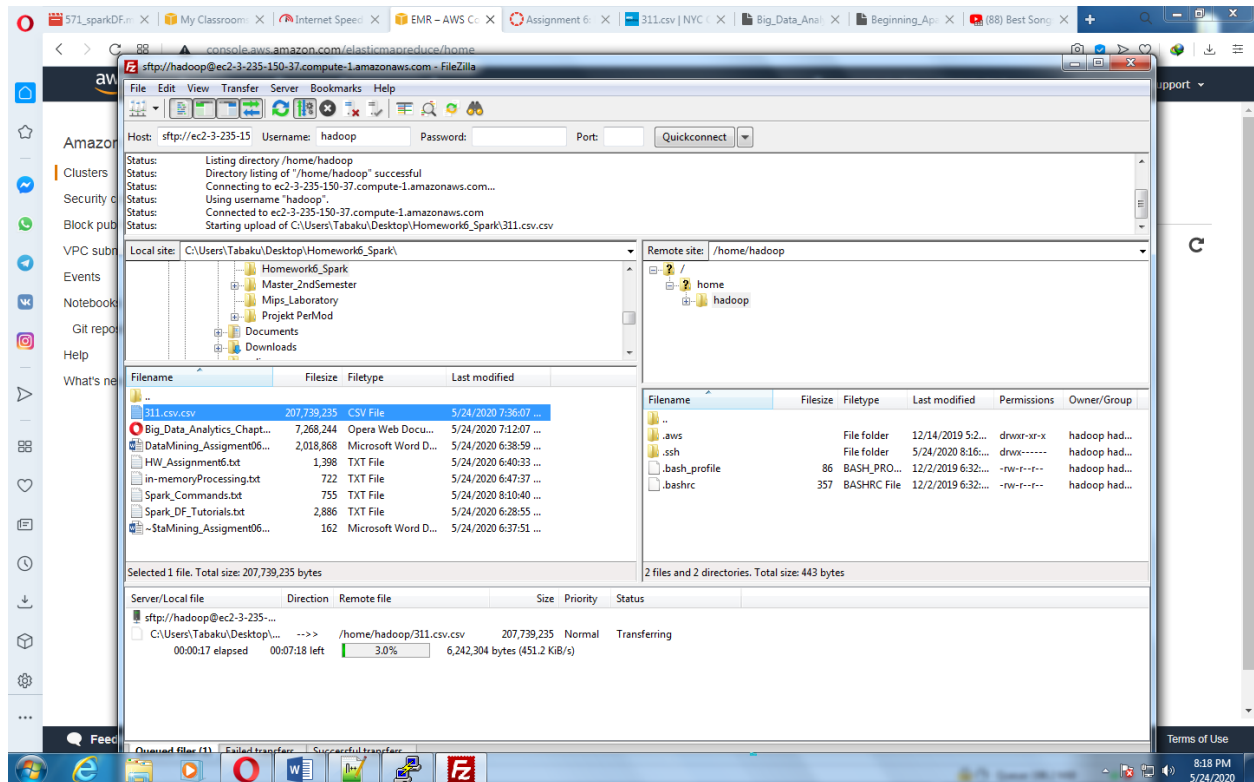
ACCEPTED:

Prof.Dr. Arben Asllani

Tasks to complete Use 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. Starting the spark shell and proceeding further as shown below

```

Spark session available as 'spark'.
Welcome to

  ____
 /_  __ \
/_/_/  \_\_
      \_\_\_

version 2.4.4

Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_252)
Type in expressions to have them evaluated.
Type :help for more information.

scala>

scala> val d311DataSet = spark.read.format("csv").option("header", "true").load("/SPRK/Ex2/311.csv")
d311DataSet: org.apache.spark.sql.DataFrame = [Unique Key: string, Created Date: string ... 36 more fields]

scala> d311DataSet.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)

```

```
scala> d311DataSet.show()
20/05/24 19:16:38 WARN Utils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.debug.maxToStringFields' in SparkEnv.conf.
```

[Unique Key]	Created Date	Closed Date	Agency	Agency Name	Complaint Type	Descriptor
ocation Type Incident Zip Incident Address Street Name Cross Street 1 Cross Street 2 Intersection Street 1 Inter	ection 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 Park Borough Vehicle Type Taxi Co	mpany Borough Taxi Pick Up Location Bridge Highway Name Bridge Highway Direction Road Ramp Bridge Highway Segment Latitude	Longitude Location		
28163399 06/01/2014 12:00:... 06/01/2014 12:15:... 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 06/01/2014 12:15:... Unspe	cified BROOKLYN BROOKLYN null null null Unspecified BROOKLYN null null null			
28157590 06/01/2014 12:00:... 06/06/2014 04:03:... DOHMH Department of Hea... Rodent Mouse Sighting 1-2 Fam	ily Dwelling 11372 170-04 ROOSEVELT A... ROOSEVELT AVENUE 70 STREET BQE WESTBOUND ENT... null	null ADDRESS Jackson Heights null N/A Closed 07/01/2014 12:45:... 06/06/2014 04:03:...	02 QUEENS QUEENS 1013248 211238 Unspecified QUEENS null null 40.7464285492033			
28157974 06/01/2014 12:00:... 06/10/2014 12:00:... HPD Department of Hou... UNSANITARY CONDITION GARBAGE/RECYCLING... RESIDENT	IAL BUILDING 10024 336 WEST 77 STREET WEST 77 STREET WEST END AVENUE RIVERSIDE DRIVE null	null ADDRESS NEW YORK null N/A Closed null 06/10/2014 12:00:...	07 MANHATTAN MANHATTAN 988998 224663 Unspecified MANHATTAN null null 40.78332297541809			
28158733 06/01/2014 12:00:... 06/09/2014 12:00:... HPD Department of Hou... UNSANITARY CONDITION PESTS RESIDENT	IAL BUILDING 11355 140-37 ASH AVENUE ASH AVENUE KISSENA BOULEVARD BOWNE STREET null	null ADDRESS Flushing null N/A Closed null 06/09/2014 12:00:...	07 QUEENS QUEENS 1033351 214635 Unspecified QUEENS null null 40.75566366565145			
28159418 06/01/2014 12:00:... 06/20/2014 12:00:... HPD Department of Hou... UNSANITARY CONDITION MOLD RESIDENT	IAL BUILDING 10458 2604 BAINBRIDGE A... BAINBRIDGE AVENUE EAST 193 STREET EAST 194 STREET null					

```
scala> d311DataSet.show(5)
-----
[Unique Key]      Created Date|      Closed Date|Agency|      Agency Name|      Complaint Type|      Descriptor|      Location Type|Incident Zip|      Incident Address
a|      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|Park Borough|Vehicle Type|Taxi Comp
ny Borough|Taxi Pick Up Location|Bridge Highway Name|Bridge Highway Direction|Road Ramp|Bridge Highway Segment|      Latitude|      Longitude|      Location|
-----
| 2816399|06/01/2014 12:00:...|06/01/2014 12:15:...|DOT|Department of Tra...|Traffic Signal Co...|LED Pedestrian Unit|      null|      null|      null|      null|
a|      null|      null|      null|      18 AVE|      4 ST E|INTERSECTION|      null|      null|      null|N/A|Closed|      null|
null|      06/01/2014 12:15:...|Unspecified BROOKLYN|BROOKLYN|      null|      null|      null|      null|      null|Unspecified|      BROOKLYN|      null|
| 2815790|06/01/2014 12:00:...|06/06/2014 04:03:...|DOHMH|Department of Hea...|      null|      null|      null|      null|      null|Unspecified|      null|      null|
|ROOSEVELT AVENUE|      70 STREET|BQE WESTBOUND ENT...|      null|      null|      null|      null|      null|Unspecified|      QUEENS|      null|
91...|      06/06/2014 04:03:...|      02 QUEENS|QUEENS|      null|      null|      null|      null|      null|Unspecified|      QUEENS|      null|
| 2815794|06/01/2014 12:00:...|06/10/2014 12:00:...|HPD|Department of Hou...|UNSANITARY CONDITION|GARBAGE/RECYCLING...|RESIDENTIAL BUILDING|      null|      null|      null|      null|
T|WEST 77 STREET|WEST END AVENUE|      RIVERSIDE DRIVE|      null|      null|      null|      null|      null|Unspecified|      null|      null|
|      06/10/2014 12:00:...|      07 MANHATTAN|MANHATTAN|      null|      null|      null|      null|      null|Unspecified|      MANHATTAN|      null|
| 28158733|06/01/2014 12:00:...|06/09/2014 12:00:...|HPD|Department of Hou...|UNSANITARY CONDITION|      null|      null|      null|      null|      null|Unspecified|      QUEENS|      null|
E|      ASH AVENUE|KISSENA BOULEVARD|      BOWNE STREET|      null|      null|      null|      null|      null|Unspecified|      null|      null|
|      null|      06/09/2014 12:00:...|      07 QUEENS|QUEENS|      null|      null|      null|      null|      null|Unspecified|      QUEENS|      null| | |
| 28159418|06/01/2014 12:00:...|06/20/2014 12:00:...|HPD|Department of Hou...|UNSANITARY CONDITION|      null|      null|      null|      null|      null|Unspecified|      null|      null|
|BAINBRIDGE AVENUE|EAST 194 STREET|EAST 194 STREET|      null|      null|      null|      null|      null|Unspecified|      null|      null|
|      null|      06/20/2014 12:00:...|      null|BRONX|BRONX|      null|      null|      null|      null|      null|Unspecified|      null|      null|
|      null|      null|      null|      null|      null|      null|      null|      null|      null|Unspecified|      null|      null|
-----
only showing top 5 rows

scala>
```

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> █
```