CLEANING OF DATA

1. The original data set contained 201404 rows and 28 columns.
2. After cleaning the data, only 6 columns are required for our analysis
3. Final output is saved in the folder ‘cleandata’ on HDFS and 2 part files are formed.

```
sa5476@login-2-1:~

scala> finalOutput.saveAsTextFile("/user/sa5476/bdad/project/cleandata")

scala> finalOutput.take(5).foreach(println)
"DATA_YEAR","STATE_NAME","JUVENILE_VICTIM_COUNT","VICTIM_COUNT","LOCATION_NAME",
"MULTIPLE_OFFENSE"
1991,"Arkansas",31-AUG-91,"White","Intimidation","Anti-Black or African American
"
1991,"Arkansas",19-SEP-91,"Black or African American","Simple Assault","Anti-Whi
te"
1991,"Arkansas",04-JUL-91,"Black or African American","Aggravated Assault","Anti
-Black or African American"
1991,"Arkansas",24-DEC-91,"Black or African American","Aggravated Assault;Destru
ction/Damage/Vandalism of Property","Anti-White"

scala>
```

```
[sa5476@login-2-1 ~]$ hdfs dfs -ls /user/sa5476/bdad/project/cleandata
Found 3 items
-rw-r--r--+  3 sa5476 users          0 2020-04-11 20:24 /user/sa5476/bdad/projec
t/cleandata/_SUCCESS
-rw-r--r--+  3 sa5476 users    7980290 2020-04-11 20:24 /user/sa5476/bdad/projec
t/cleandata/part-00000
-rw-r--r--+  3 sa5476 users    8014132 2020-04-11 20:24 /user/sa5476/bdad/projec
t/cleandata/part-00001
[sa5476@login-2-1 ~]$
```

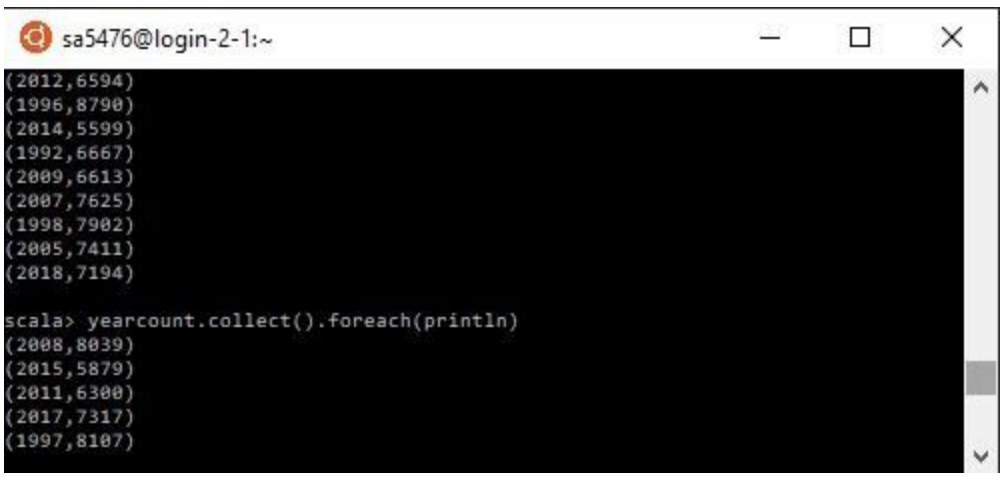
Code Snippet for Cleaning the Data

```scala
val crimedata = "/user/sa5476/bdad/project/hate_crime.csv"

val crime = sc.textFile(crimedata)

val output = crime.map(line => line.split(","))

val validCols = output.map(array => (array(1),array(7),array(14),array(20),array(23),array(26)))

val finalOutput = validCols.map { a => a.productIterator.mkString(",") }

finalOutput.saveAsTextFile("/user/sa5476/bdad/project/cleandata")
```

PROFILING OF DATA

1. The cleaned dataset is used further profiling by counting the crimes that occurred each year.
2. This profiling is done by using reduceByKey where the key of the crime is year during which it happened.
3. Counting this year count gives 28 which is equal to the total number of given years in the dataset.
4. This data will be further used to determine the crime and type o crime each year.

5. Here, State, Offender_Race, Offence_Name, Bias_Description, Date will be of string type; whereas Year is of int type.
6. The number of columns in the final dataset is 6 and 201404 rows.



Code Snippet for Profiling the Data

```
val crimedata = "/user/sa5476/bdad/project/hate_crime.csv"

val crime = sc.textFile(crimedata)

val output = crime.map(line => line.split(","))

val yeardata = output.map(array => (array(1),1))

val yearcount = yeardata.reduceByKey((x,y) => x+y)

yearcount.count() //Long = 28
```