Lab Assignment 2

TEAM ID : 10

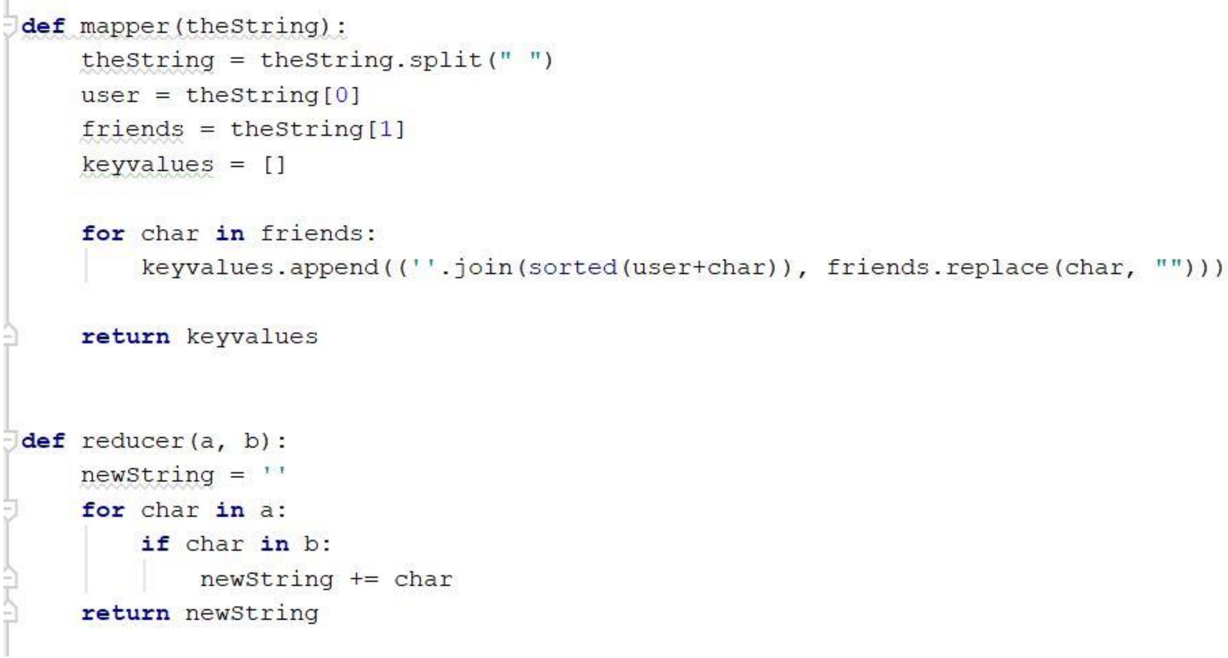
Team Members :

- Nikhil Shravan Krishna Sanka (23)
- Tejaswi Ayyadapu (2)
- Sumanth Sanakkayala (22)

Part 1

Aim :

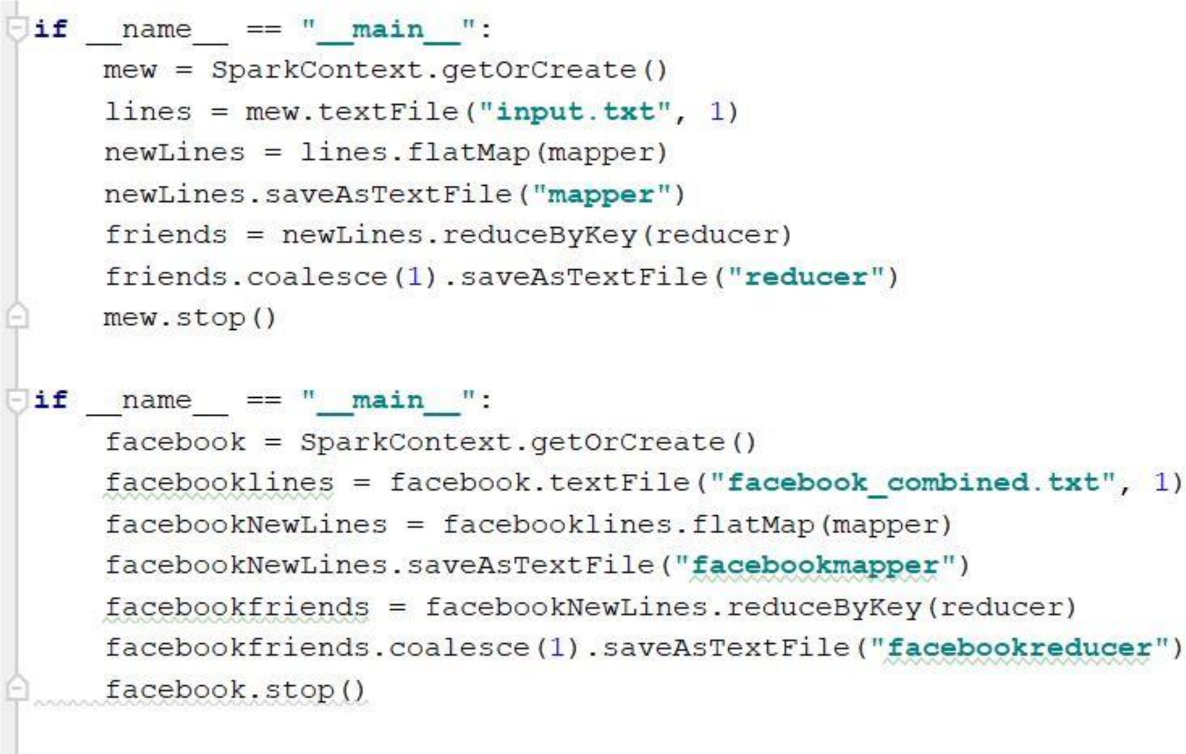
Implement MapReduce algorithm for finding Facebook common friends problem and run the MapReduce job on Apache Spark.



```python
def mapper(theString):
    theString = theString.split(" ")
    user = theString[0]
    friends = theString[1]
    keyvalues = []

    for char in friends:
        keyvalues.append((''.join(sorted(user+char)), friends.replace(char, "")))

    return keyvalues


def reducer(a, b):
    newString = ''
    for char in a:
        if char in b:
            newString += char
    return newString
```

I have used two main functions :

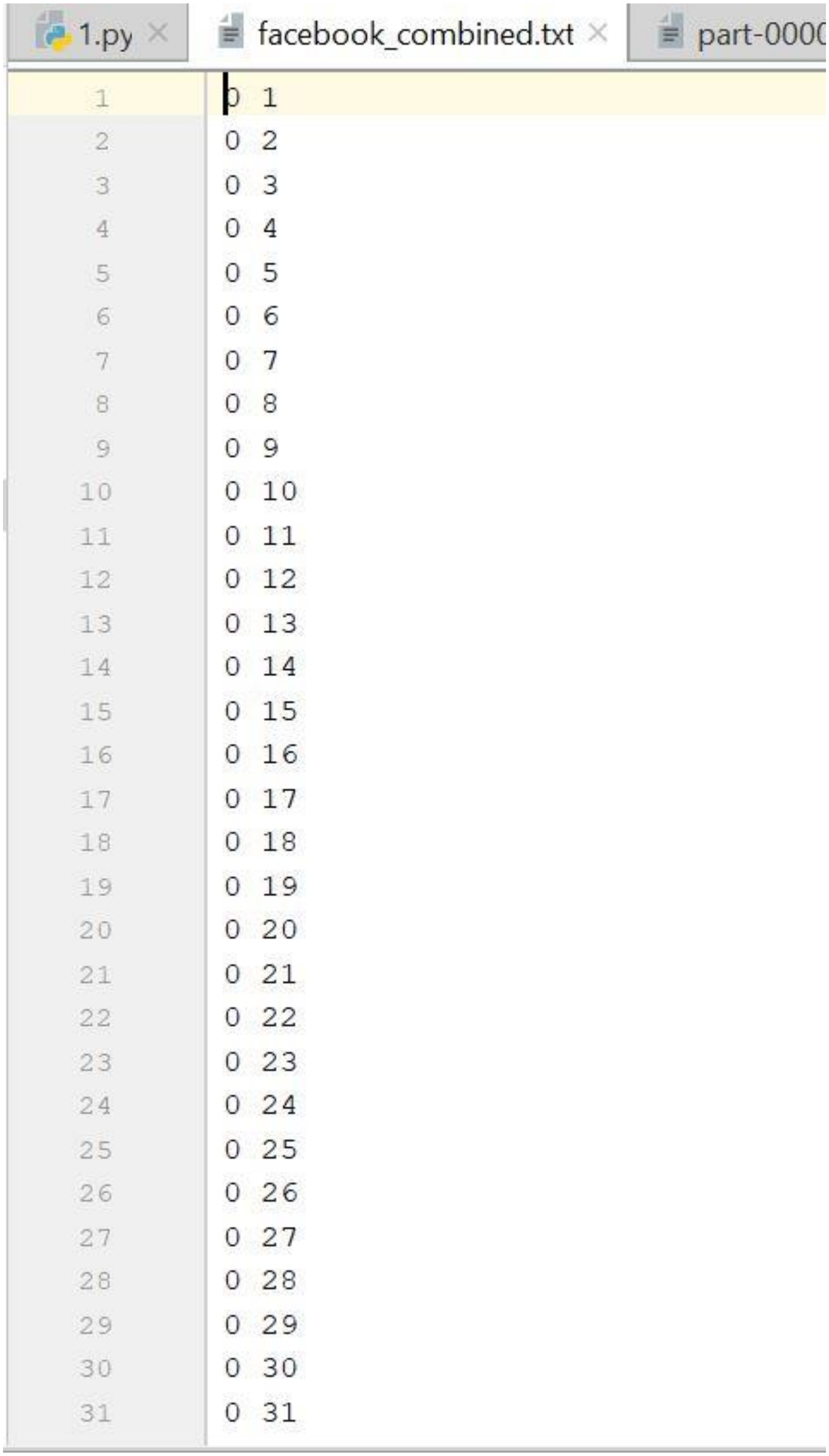
- One for the dataset given.
- Other for the simple dataset.



```
if __name__ == "__main__":
    mew = SparkContext.getOrCreate()
    lines = mew.textFile("input.txt", 1)
    newLines = lines.flatMap(mapper)
    newLines.saveAsTextFile("mapper")
    friends = newLines.reduceByKey(reducer)
    friends.coalesce(1).saveAsTextFile("reducer")
    mew.stop()

if __name__ == "__main__":
    facebook = SparkContext.getOrCreate()
    facebooklines = facebook.textFile("facebook_combined.txt", 1)
    facebookNewLines = facebooklines.flatMap(mapper)
    facebookNewLines.saveAsTextFile("facebookmapper")
    facebookfriends = facebookNewLines.reduceByKey(reducer)
    facebookfriends.coalesce(1).saveAsTextFile("facebookreducer")
    facebook.stop()
```

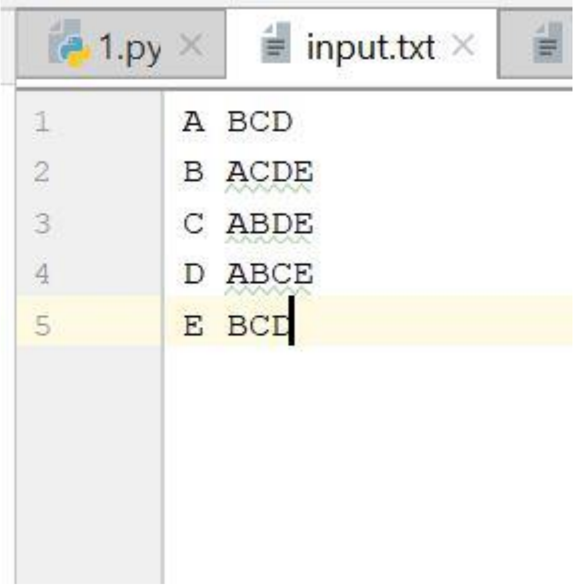
Input :

- Using the given dataset.



```
0 1
0 2
0 3
0 4
0 5
0 6
0 7
0 8
0 9
0 10
0 11
0 12
0 13
0 14
0 15
0 16
0 17
0 18
0 19
0 20
0 21
0 22
0 23
0 24
0 25
0 26
0 27
0 28
0 29
0 30
0 31
```

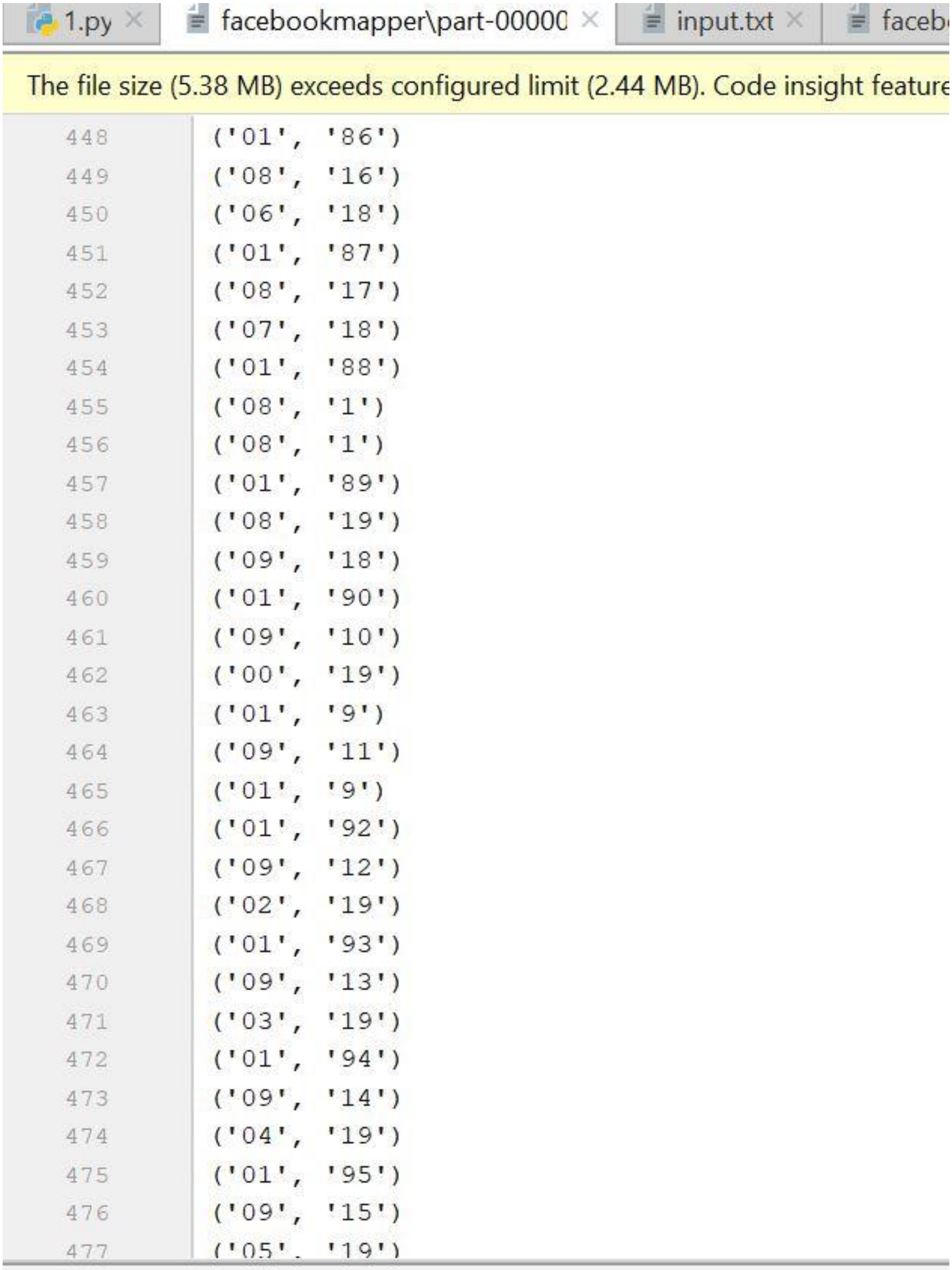
- Using the simple dataset.



```
A BCD
B ACDE
C ABDE
D ABCE
E BCD
```

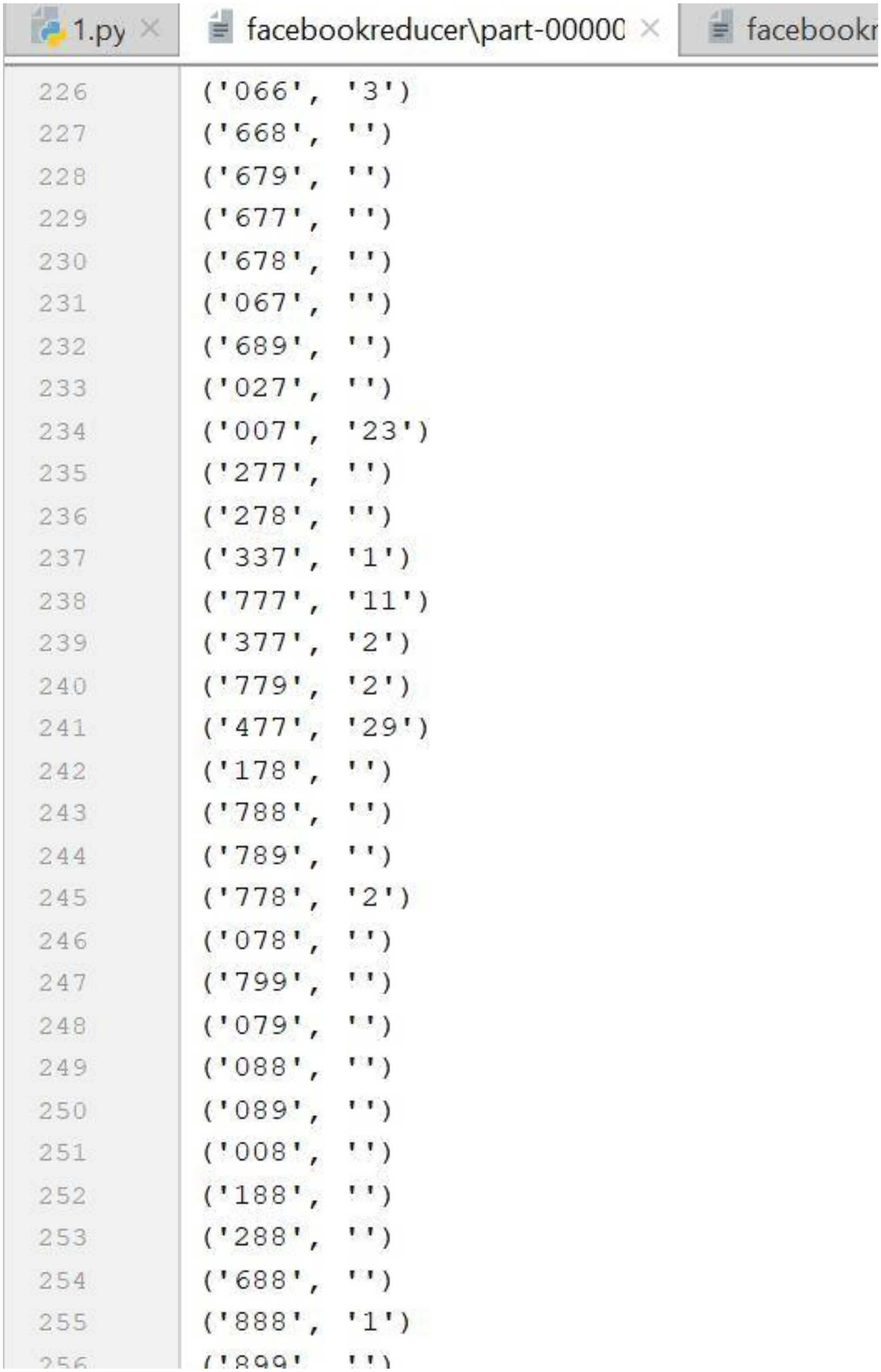
Output :

- For the given dataset.



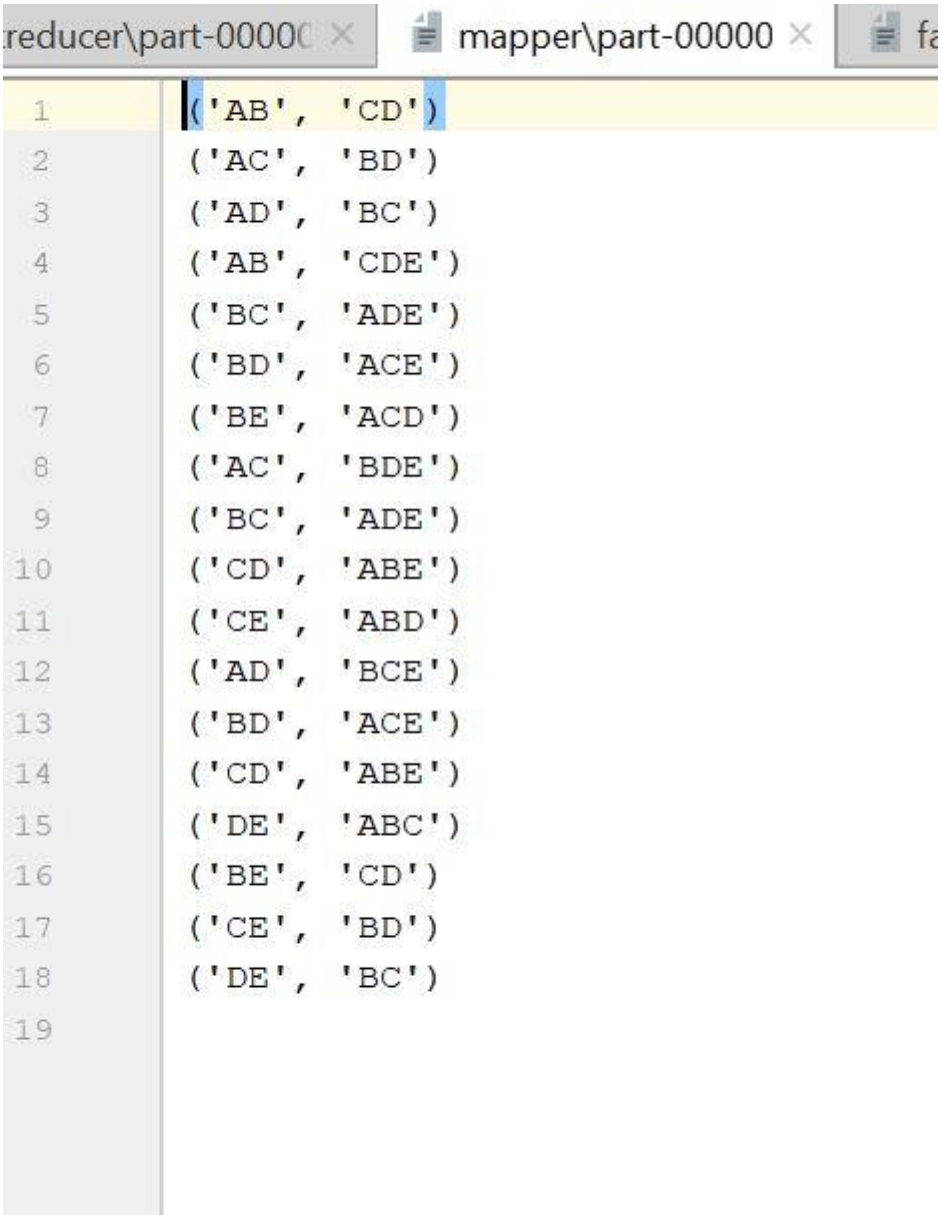
The file size (5.38 MB) exceeds configured limit (2.44 MB). Code insight feature

```
('01', '86')
('08', '16')
('06', '18')
('01', '87')
('08', '17')
('07', '18')
('01', '88')
('08', '1')
('08', '1')
('01', '89')
('08', '19')
('09', '18')
('01', '90')
('09', '10')
('00', '19')
('01', '9')
('09', '11')
('01', '9')
('01', '92')
('09', '12')
('02', '19')
('01', '93')
('09', '13')
('03', '19')
('01', '94')
('09', '14')
('04', '19')
('01', '95')
('09', '15')
('05', '19')
```

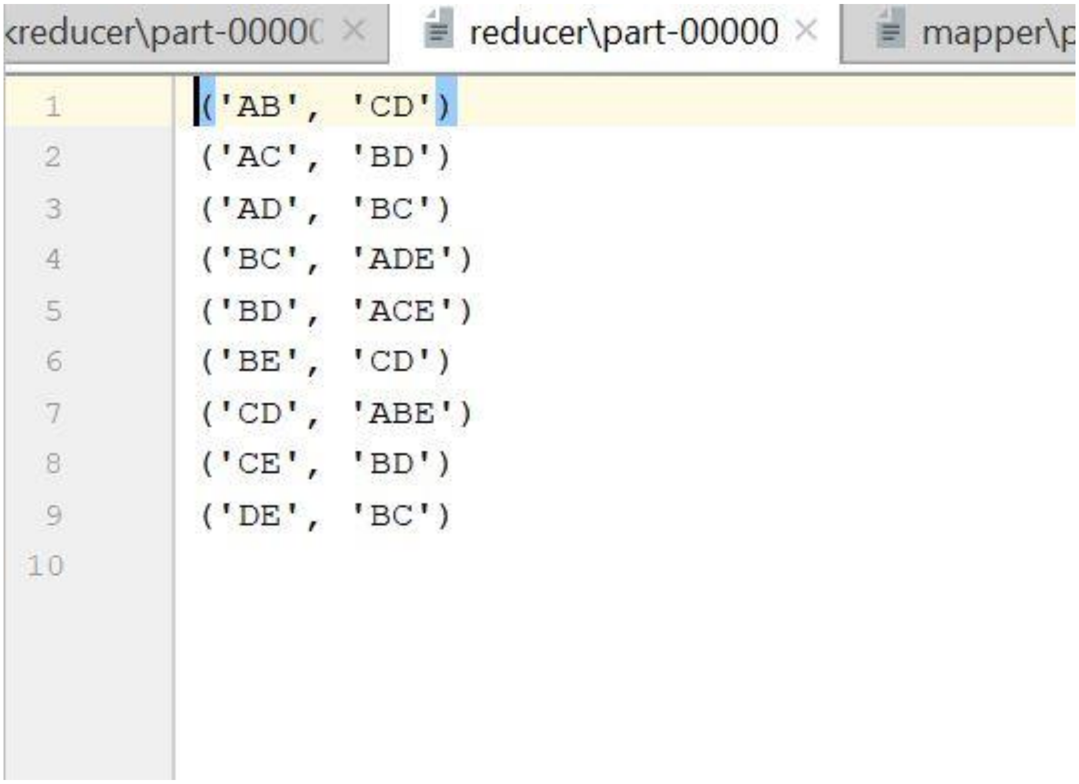


```
('066', '3')
('668', '')
('679', '')
('677', '')
('678', '')
('067', '')
('689', '')
('027', '')
('007', '23')
('277', '')
('278', '')
('337', '1')
('777', '11')
('377', '2')
('779', '2')
('477', '29')
('178', '')
('788', '')
('789', '')
('778', '2')
('078', '')
('799', '')
('079', '')
('088', '')
('089', '')
('008', '')
('188', '')
('288', '')
('688', '')
('888', '1')
('899', '')
```

- For the simple dataset.



```
('AB', 'CD')
('AC', 'BD')
('AD', 'BC')
('AB', 'CDE')
('BC', 'ADE')
('BD', 'ACE')
('BE', 'ACD')
('AC', 'BDE')
('BC', 'ADE')
('CD', 'ABE')
('CE', 'ABD')
('AD', 'BCE')
('BD', 'ACE')
('CD', 'ABE')
('DE', 'ABC')
('BE', 'CD')
('CE', 'BD')
('DE', 'BC')
```



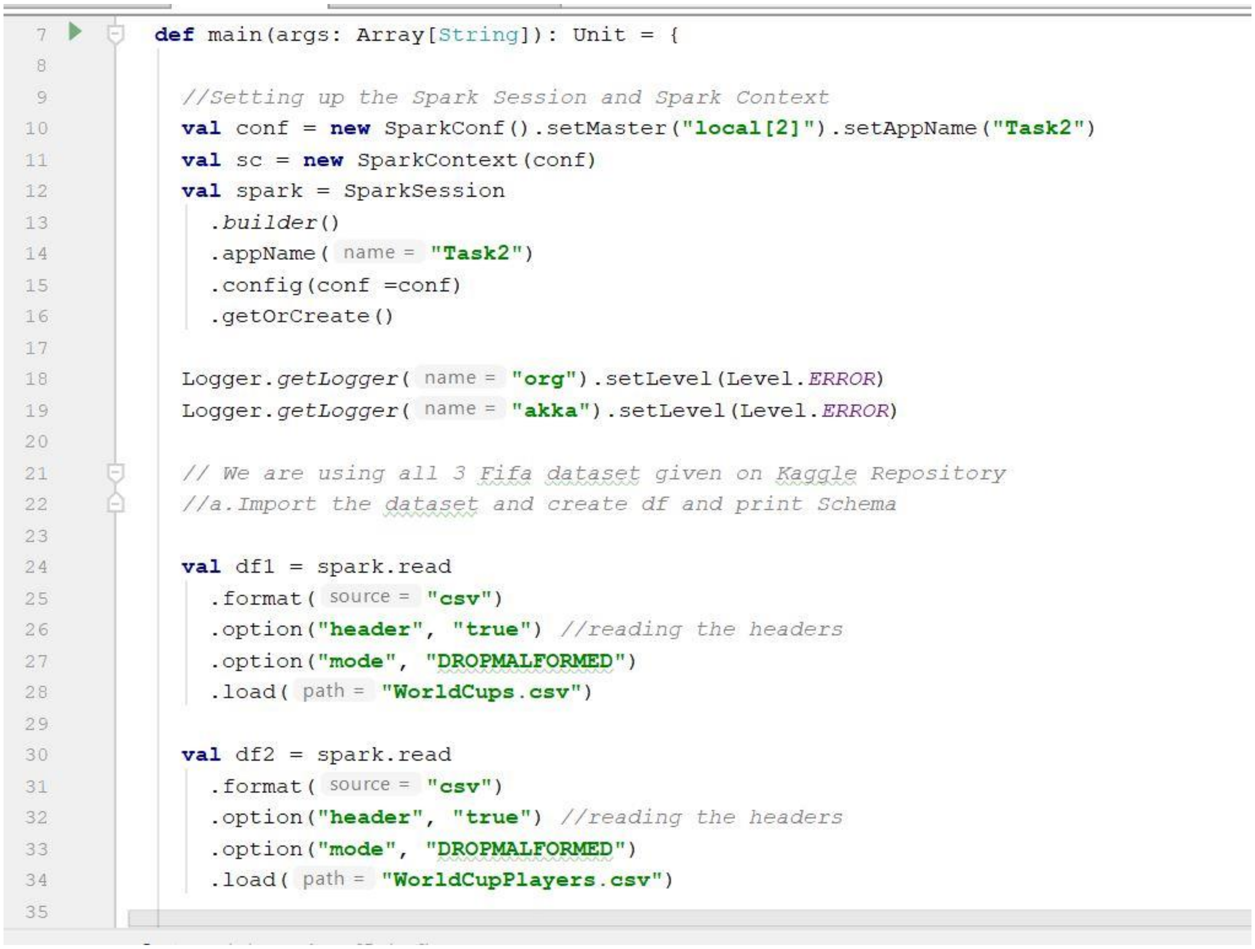
```
('AB', 'CD')
('AC', 'BD')
('AD', 'BC')
('BC', 'ADE')
('BD', 'ACE')
('BE', 'CD')
('CD', 'ABE')
('CE', 'BD')
('DE', 'BC')
```

Part 2

Aim :

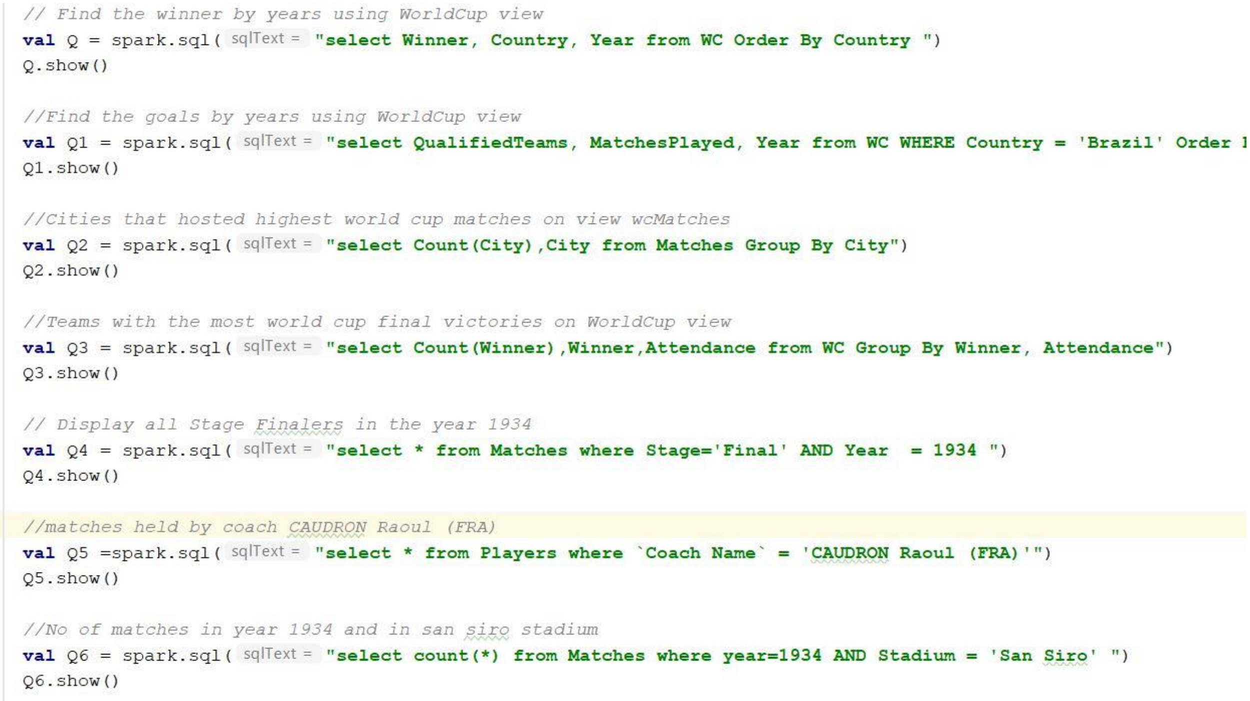
1. Create a Spark DataFrame using one of datasets and try to use all different StructType.

We have used the dataset "FIFA World Cup" for this part.

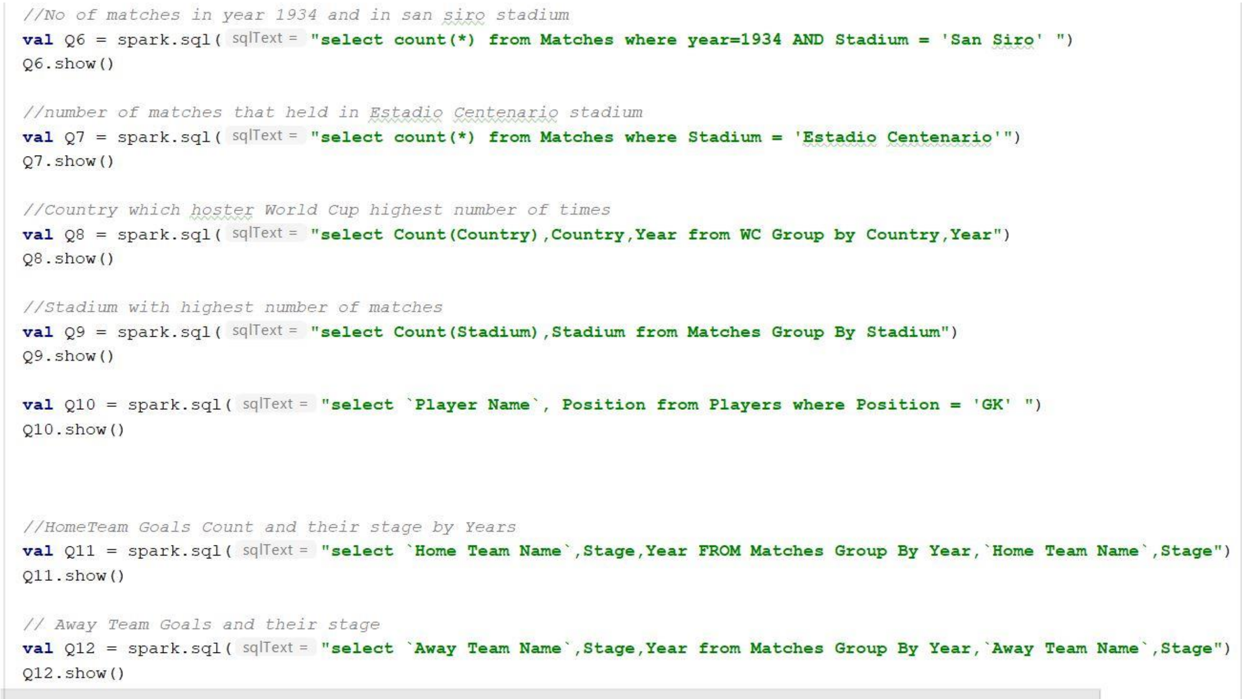


```
def main(args: Array[String]): Unit = {

  //Setting up the Spark Session and Spark Context
  val conf = new SparkConf().setMaster("local[2]").setAppName("Task2")
  val sc = new SparkContext(conf)
  val spark = SparkSession
    .builder()
    .appName( name = "Task2")
    .config(conf =conf)
    .getOrCreate()

  Logger.getLogger( name = "org").setLevel(Level.ERROR)
  Logger.getLogger( name = "akka").setLevel(Level.ERROR)

  // We are using all 3 Fifa dataset given on Kaggle Repository
  //a.Import the dataset and create df and print Schema

  val df1 = spark.read
    .format( source = "csv")
    .option("header", "true") //reading the headers
    .option("mode", "DROPMALFORMED")
    .load( path = "WorldCups.csv")

  val df2 = spark.read
    .format( source = "csv")
    .option("header", "true") //reading the headers
    .option("mode", "DROPMALFORMED")
    .load( path = "WorldCupPlayers.csv")

```

2. Perform 10 intuitive questions in Dataset.

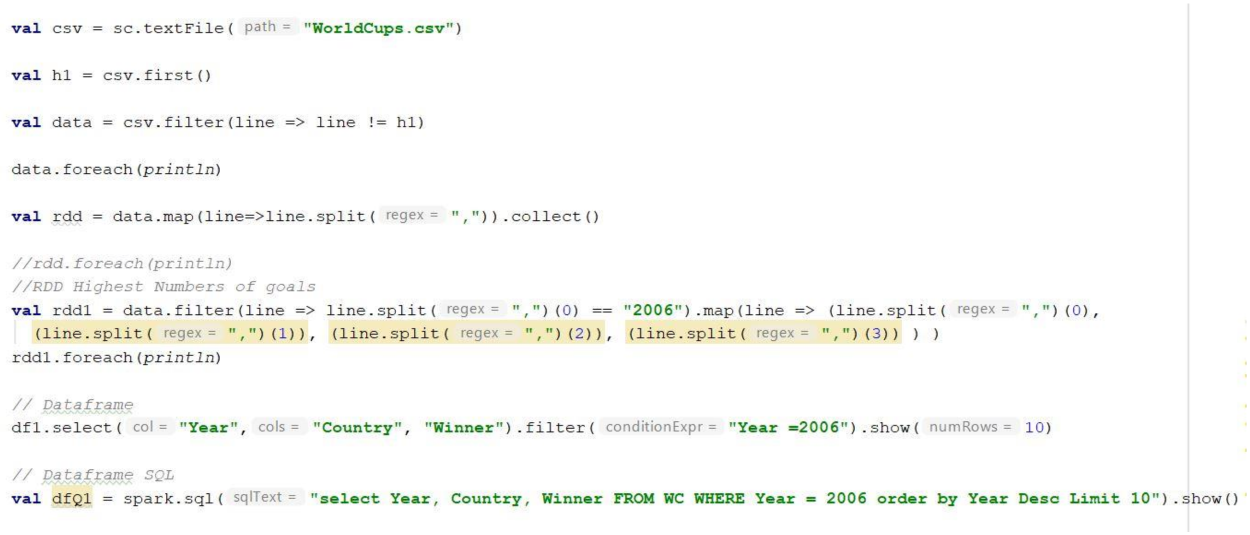


```
// Find the winner by years using WorldCup view
val Q = spark.sql( sqlText = "select Winner, Country, Year from WC Order By Country ")
Q.show()

//Find the goals by years using WorldCup view
val Q1 = spark.sql( sqlText = "select QualifiedTeams, MatchesPlayed, Year from WC WHERE Country = 'Brazil' Order B
Q1.show()

//Cities that hosted highest world cup matches on view wcMatches
val Q2 = spark.sql( sqlText = "select Count(City),City from Matches Group By City")
Q2.show()

//Teams with the most world cup final victories on WorldCup view
val Q3 = spark.sql( sqlText = "select Count(Winner),Winner,Attendance from WC Group By Winner, Attendance")
Q3.show()

// Display all Stage Finalers in the year 1934
val Q4 = spark.sql( sqlText = "select * from Matches where Stage='Final' AND Year  = 1934 ")
Q4.show()

//matches held by coach CAUDRON Raoul (FRA)
val Q5 =spark.sql( sqlText = "select * from Players where `Coach Name` = 'CAUDRON Raoul (FRA)'")
Q5.show()

//No of matches in year 1934 and in san siro stadium
val Q6 = spark.sql( sqlText = "select count(*) from Matches where year=1934 AND Stadium = 'San Siro' ")
Q6.show()
```

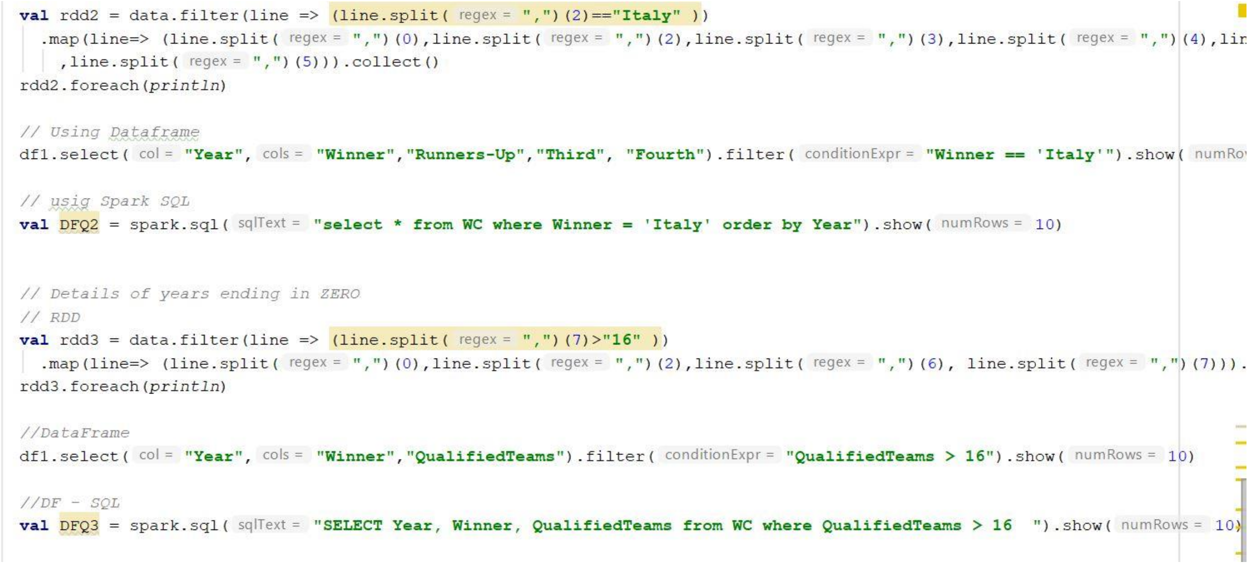


```
//No of matches in year 1934 and in san siro stadium
val Q6 = spark.sql( sqlText = "select count(*) from Matches where year=1934 AND Stadium = 'San Siro' ")
Q6.show()

//number of matches that held in Estadio Centenario stadium
val Q7 = spark.sql( sqlText = "select count(*) from Matches where Stadium = 'Estadio Centenario'")
Q7.show()

//Country which hoster World Cup highest number of times
val Q8 = spark.sql( sqlText = "select Count(Country),Country,Year from WC Group by Country,Year")
Q8.show()

//Stadium with highest number of matches
val Q9 = spark.sql( sqlText = "select Count(Stadium),Stadium from Matches Group By Stadium")
Q9.show()

val Q10 = spark.sql( sqlText = "select `Player Name`, Position from Players where Position = 'GK' ")
Q10.show()



//HomeTeam Goals Count and their stage by Years
val Q11 = spark.sql( sqlText = "select `Home Team Name`,Stage,Year FROM Matches Group By Year,`Home Team Name`,Stage")
Q11.show()

// Away Team Goals and their stage
val Q12 = spark.sql( sqlText = "select `Away Team Name`,Stage,Year from Matches Group By Year,`Away Team Name`,Stage")
Q12.show()
```

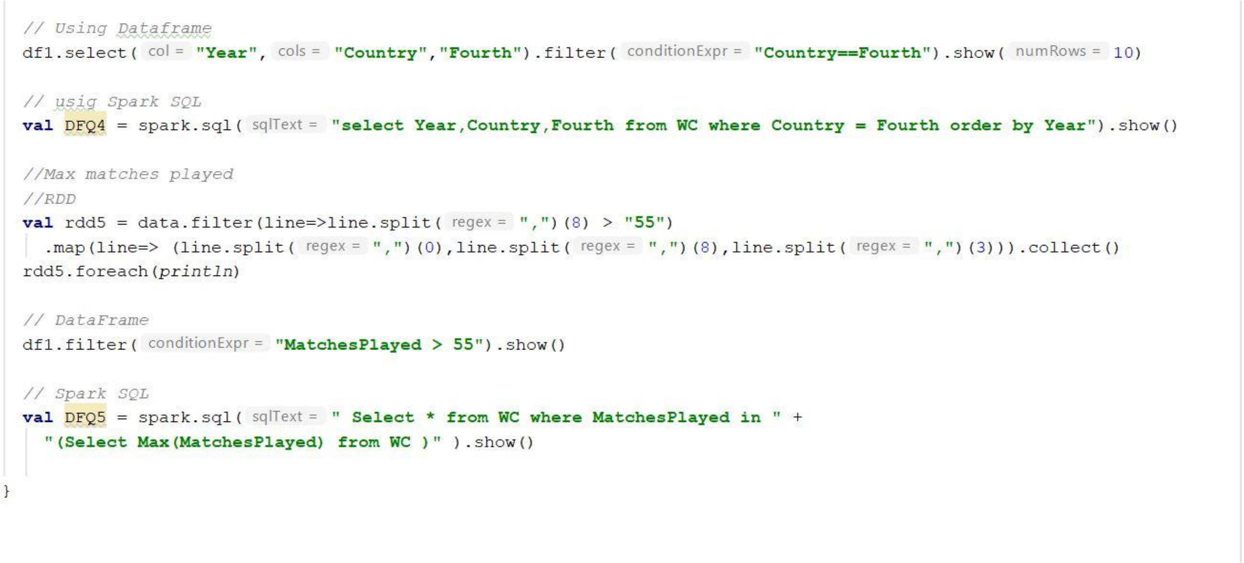
3. Perform any 5 queries in Spark RDD’s and Spark Data Frames.



```
val csv = sc.textFile( path = "WorldCups.csv")

val h1 = csv.first()

val data = csv.filter(line => line != h1)

data.foreach(println)

val rdd = data.map(line=>line.split( regex = ",")).collect()

//rdd.foreach(println)
//RDD Highest Numbers of goals
val rdd1 = data.filter(line => line.split( regex = ",")(0) == "2006").map(line => (line.split( regex = ",")(0),
  (line.split( regex = ",")(1)), (line.split( regex = ",")(2)), (line.split( regex = ",")(3)) ) )
rdd1.foreach(println)

// Dataframe
df1.select( col = "Year", cols = "Country", "Winner").filter( conditionExpr = "Year =2006").show( numRows = 10)

// Dataframe SQL
val dfQ1 = spark.sql( sqlText = "select Year, Country, Winner FROM WC WHERE Year = 2006 order by Year Desc Limit 10").show()
```

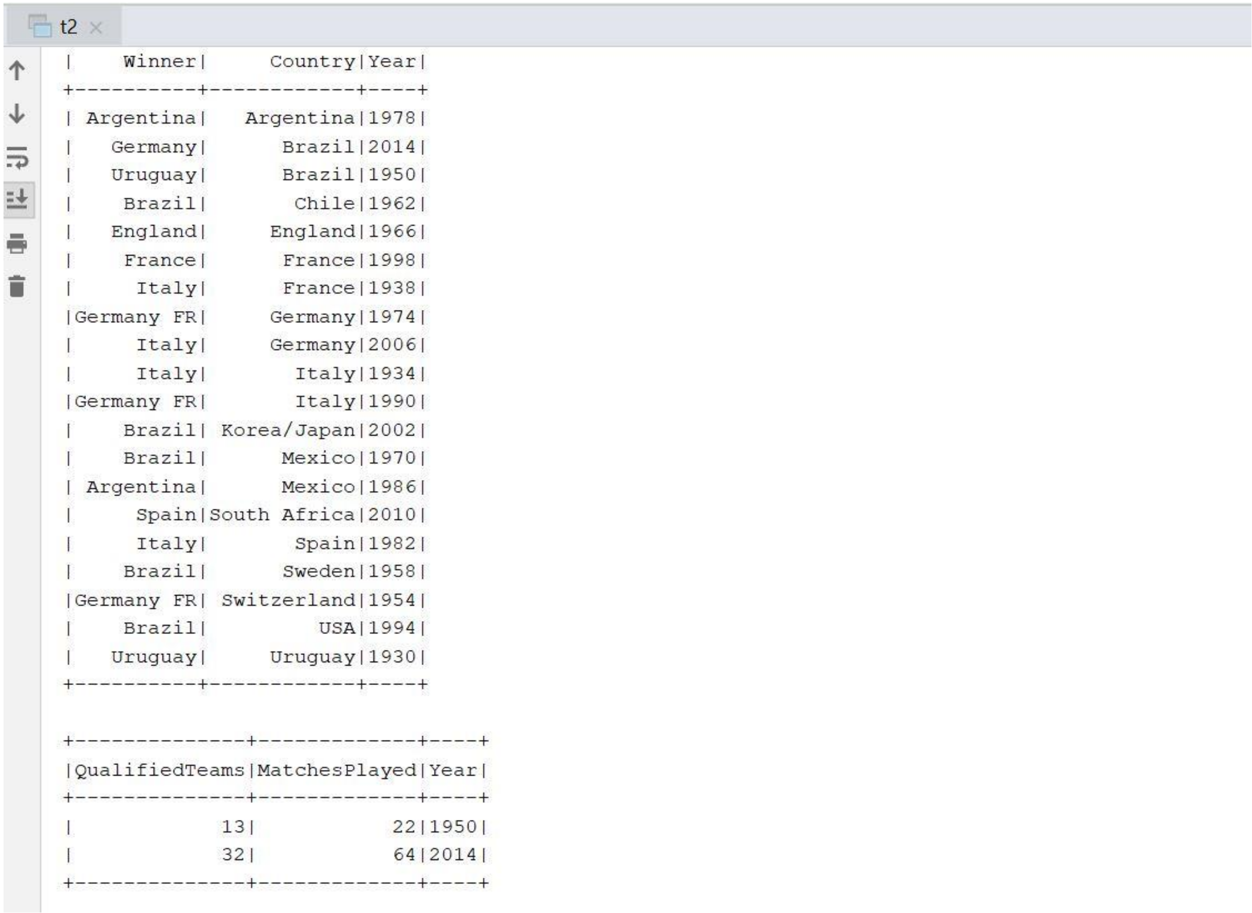


```
val rdd2 = data.filter(line => (line.split( regex = ",")(2)=="Italy" ))
  .map(line=> (line.split( regex = ",")(0),line.split( regex = ",")(2),line.split( regex = ",")(3),line.split( regex = ",")(4),lin
    ,line.split( regex = ",")(5))).collect()
rdd2.foreach(println)

// Using Dataframe
df1.select( col = "Year", cols = "Winner","Runners-Up","Third", "Fourth").filter( conditionExpr = "Winner == 'Italy'").show( numRo

// usig Spark SQL
val DFQ2 = spark.sql( sqlText = "select * from WC where Winner = 'Italy' order by Year").show( numRows = 10)


// Details of years ending in ZERO
// RDD
val rdd3 = data.filter(line => (line.split( regex = ",")(7)>"16" ))
  .map(line=> (line.split( regex = ",")(0),line.split( regex = ",")(2),line.split( regex = ",")(6), line.split( regex = ",")(7))).
rdd3.foreach(println)

//DataFrame
df1.select( col = "Year", cols = "Winner","QualifiedTeams").filter( conditionExpr = "QualifiedTeams > 16").show( numRows = 10)

//DF - SQL
val DFQ3 = spark.sql( sqlText = "SELECT Year, Winner, QualifiedTeams from WC where QualifiedTeams > 16  ").show( numRows = 10)
```

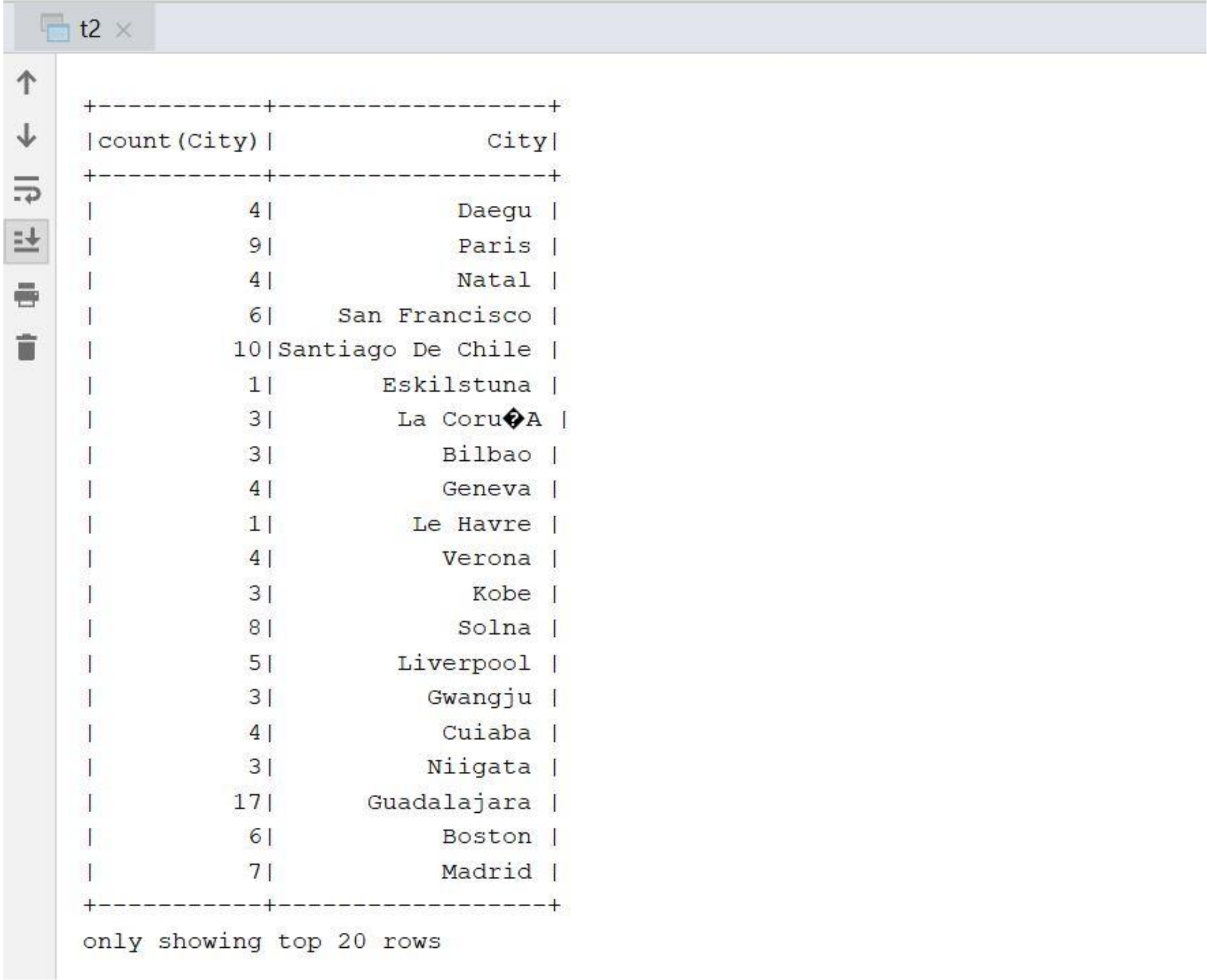


```
// Using Dataframe
df1.select( col = "Year", cols = "Country","Fourth").filter( conditionExpr = "Country==Fourth").show( numRows = 10)

// usig Spark SQL
val DFQ4 = spark.sql( sqlText = "select Year,Country,Fourth from WC where Country = Fourth order by Year").show()

//Max matches played
//RDD
val rdd5 = data.filter(line=>line.split( regex = ",")(8) > "55")
  .map(line=> (line.split( regex = ",")(0),line.split( regex = ",")(8),line.split( regex = ",")(3))).collect()
rdd5.foreach(println)

// DataFrame
df1.filter( conditionExpr = "MatchesPlayed > 55").show()

// Spark SQL
val DFQ5 = spark.sql( sqlText = " Select * from WC where MatchesPlayed in " +
  "(Select Max(MatchesPlayed) from WC )" ).show()

}
```

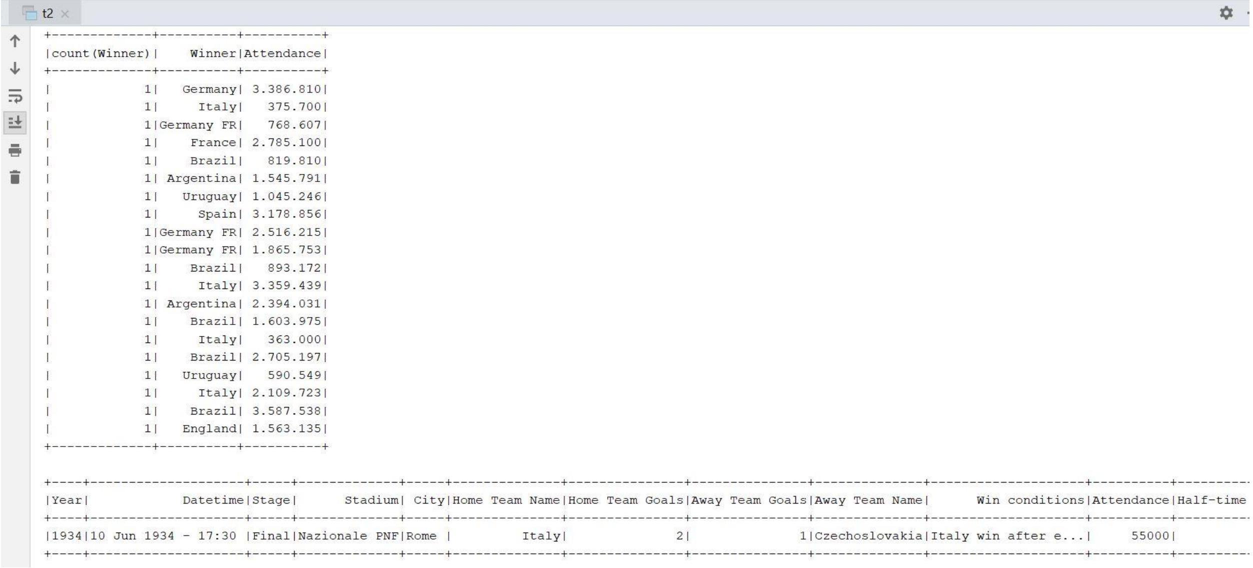
Output :



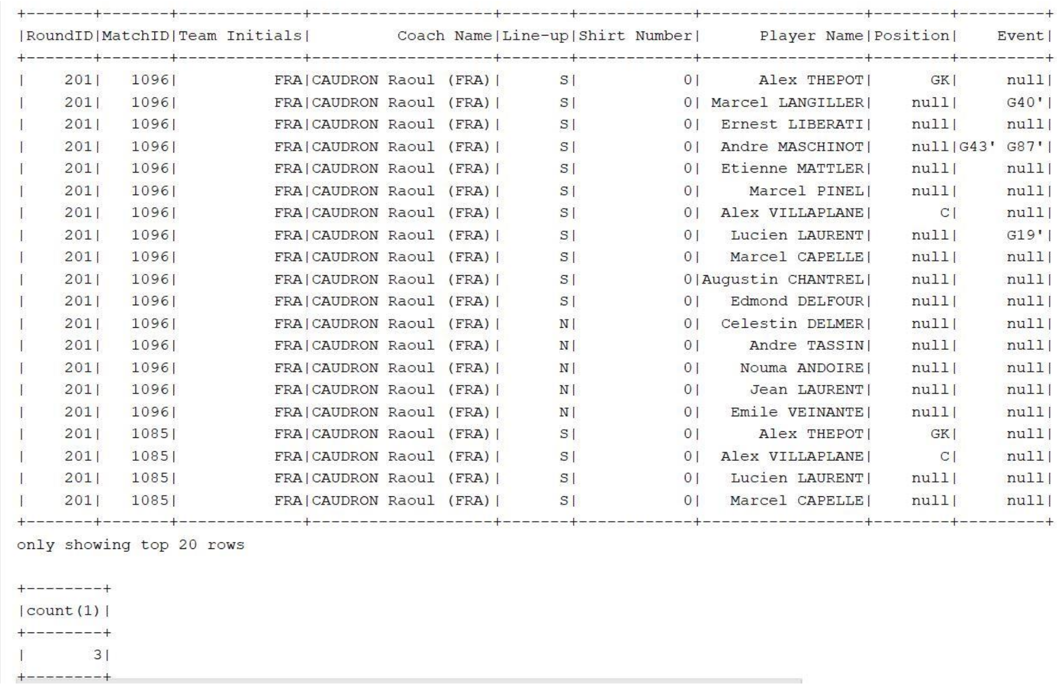
```
|    Winner|     Country|Year|
+----------+------------+----+
| Argentina|   Argentina|1978|
|   Germany|      Brazil|2014|
|   Uruguay|      Brazil|1950|
|    Brazil|       Chile|1962|
|   England|     England|1966|
|    France|      France|1998|
|     Italy|      France|1938|
|Germany FR|     Germany|1974|
|     Italy|     Germany|2006|
|     Italy|       Italy|1934|
|Germany FR|       Italy|1990|
|    Brazil| Korea/Japan|2002|
|    Brazil|      Mexico|1970|
| Argentina|      Mexico|1986|
|     Spain|South Africa|2010|
|     Italy|       Spain|1982|
|    Brazil|      Sweden|1958|
|Germany FR| Switzerland|1954|
|    Brazil|         USA|1994|
|   Uruguay|     Uruguay|1930|
+----------+------------+----+

+--------------+-------------+----+
|QualifiedTeams|MatchesPlayed|Year|
+--------------+-------------+----+
|            13|           22|1950|
|            32|           64|2014|
+--------------+-------------+----+
```



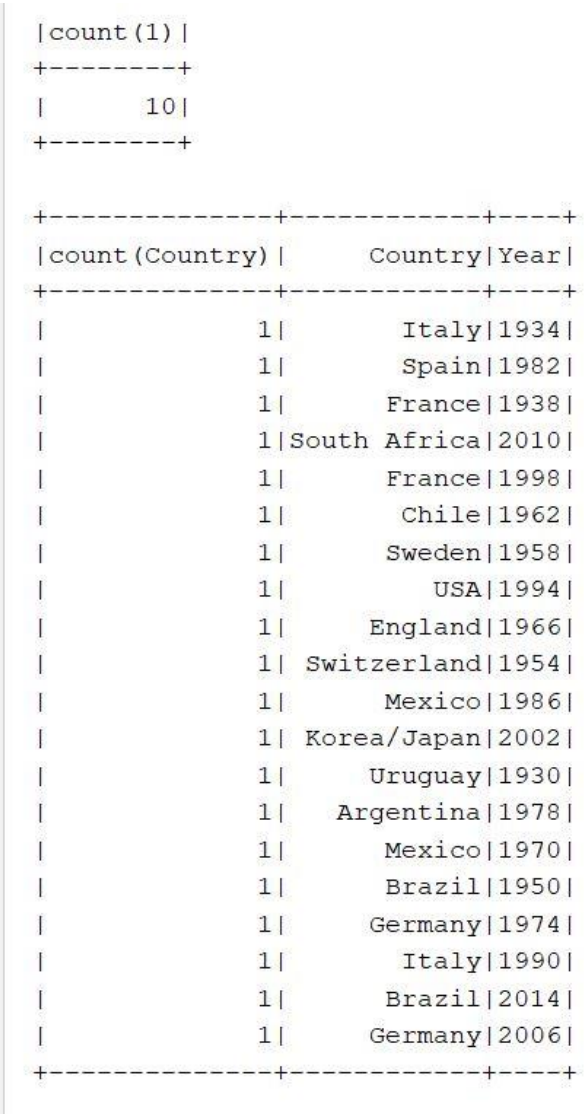
```
+-----------+------------------+
|count(City)|              City|
+-----------+------------------+
|          4|            Daegu |
|          9|            Paris |
|          4|            Natal |
|          6|    San Francisco |
|         10|Santiago De Chile |
|          1|       Eskilstuna |
|          3|         La Coru�A |
|          3|           Bilbao |
|          4|           Geneva |
|          1|         Le Havre |
|          4|           Verona |
|          3|             Kobe |
|          8|            Solna |
|          5|        Liverpool |
|          3|          Gwangju |
|          4|           Cuiaba |
|          3|          Niigata |
|         17|      Guadalajara |
|          6|           Boston |
|          7|           Madrid |
+-----------+------------------+
only showing top 20 rows
```



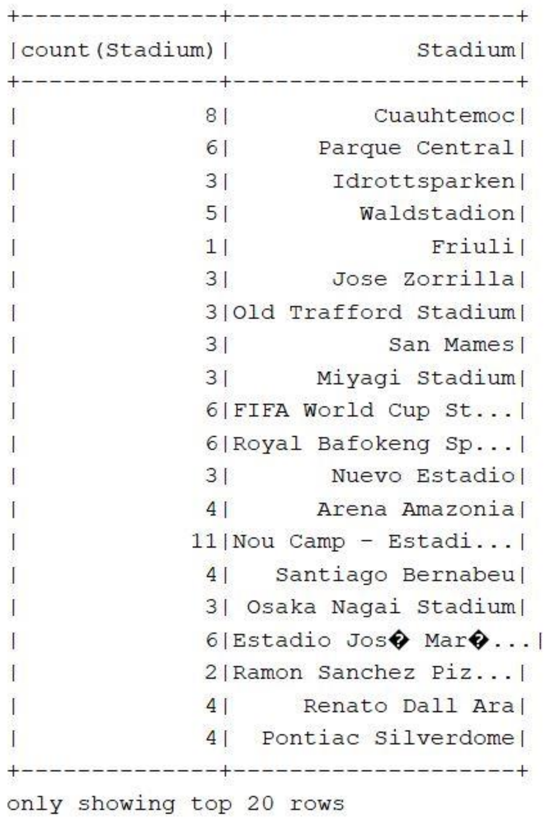
```
+-------------+----------+----------+
|count(Winner)|    Winner|Attendance|
+-------------+----------+----------+
|            1|   Germany| 3.386.810|
|            1|     Italy|   375.700|
|            1|Germany FR|   768.607|
|            1|    France| 2.785.100|
|            1|    Brazil|   819.810|
|            1| Argentina| 1.545.791|
|            1|   Uruguay| 1.045.246|
|            1|     Spain| 3.178.856|
|            1|Germany FR| 2.516.215|
|            1|Germany FR| 1.865.753|
|            1|    Brazil|   893.172|
|            1|     Italy| 3.359.439|
|            1| Argentina| 2.394.031|
|            1|    Brazil| 1.603.975|
|            1|     Italy|   363.000|
|            1|    Brazil| 2.705.197|
|            1|   Uruguay|   590.549|
|            1|     Italy| 2.109.723|
|            1|    Brazil| 3.587.538|
|            1|   England| 1.563.135|
+-------------+----------+----------+

+----+--------------------+-----+--------------+-----+--------------+---------------+---------------+---------------+--------------------+----------+----------
|Year|            Datetime|Stage|       Stadium| City|Home Team Name|Home Team Goals|Away Team Goals|Away Team Name|     Win conditions|Attendance|Half-time
+----+--------------------+-----+--------------+-----+--------------+---------------+---------------+---------------+--------------------+----------+----------
|1934|10 Jun 1934 - 17:30 |Final|Nazionale PNF|Rome |         Italy|              2|              1|Czechoslovakia|Italy win after e...|     55000|
+----+--------------------+-----+--------------+-----+--------------+---------------+---------------+---------------+--------------------+----------+----------
```



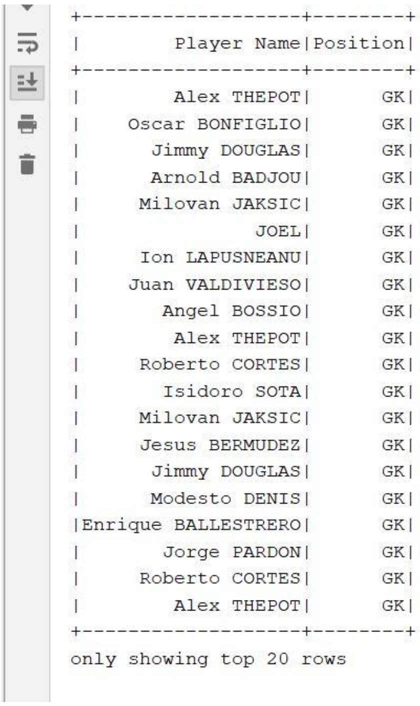
```
+-------+-------+-------------+-------------------+-------+------------+-----------------+--------+---------+
|RoundID|MatchID|Team Initials|         Coach Name|Line-up|Shirt Number|      Player Name|Position|    Event|
+-------+-------+-------------+-------------------+-------+------------+-----------------+--------+---------+
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|      Alex THEPOT|      GK|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0| Marcel LANGILLER|    null|     G40'|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Ernest LIBERATI|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Andre MASCHINOT|    null|G43' G87'|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Etienne MATTLER|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|     Marcel PINEL|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Alex VILLAPLANE|       C|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Lucien LAURENT|    null|     G19'|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Marcel CAPELLE|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|Augustin CHANTREL|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Edmond DELFOUR|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      N|           0|  Celestin DELMER|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      N|           0|     Andre TASSIN|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      N|           0|    Nouma ANDOIRE|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      N|           0|     Jean LAURENT|    null|     null|
|    201|   1096|          FRA|CAUDRON Raoul (FRA)|      N|           0|   Emile VEINANTE|    null|     null|
|    201|   1085|          FRA|CAUDRON Raoul (FRA)|      S|           0|      Alex THEPOT|      GK|     null|
|    201|   1085|          FRA|CAUDRON Raoul (FRA)|      S|           0|  Alex VILLAPLANE|       C|     null|
|    201|   1085|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Lucien LAURENT|    null|     null|
|    201|   1085|          FRA|CAUDRON Raoul (FRA)|      S|           0|   Marcel CAPELLE|    null|     null|
+-------+-------+-------------+-------------------+-------+------------+-----------------+--------+---------+
only showing top 20 rows

+--------+
|count(1)|
+--------+
|       3|
+--------+
```



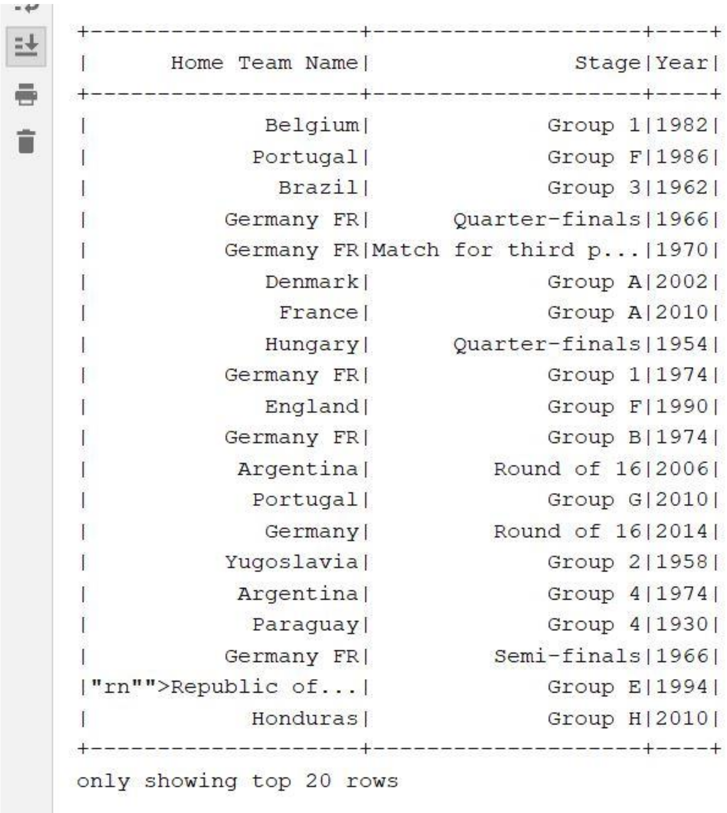
```
|count(1)|
+--------+
|      10|
+--------+

+--------------+------------+----+
|count(Country)|     Country|Year|
+--------------+------------+----+
|             1|       Italy|1934|
|             1|       Spain|1982|
|             1|      France|1938|
|             1|South Africa|2010|
|             1|      France|1998|
|             1|       Chile|1962|
|             1|      Sweden|1958|
|             1|         USA|1994|
|             1|     England|1966|
|             1| Switzerland|1954|
|             1|      Mexico|1986|
|             1| Korea/Japan|2002|
|             1|     Uruguay|1930|
|             1|   Argentina|1978|
|             1|      Mexico|1970|
|             1|      Brazil|1950|
|             1|     Germany|1974|
|             1|       Italy|1990|
|             1|      Brazil|2014|
|             1|     Germany|2006|
+--------------+------------+----+
```



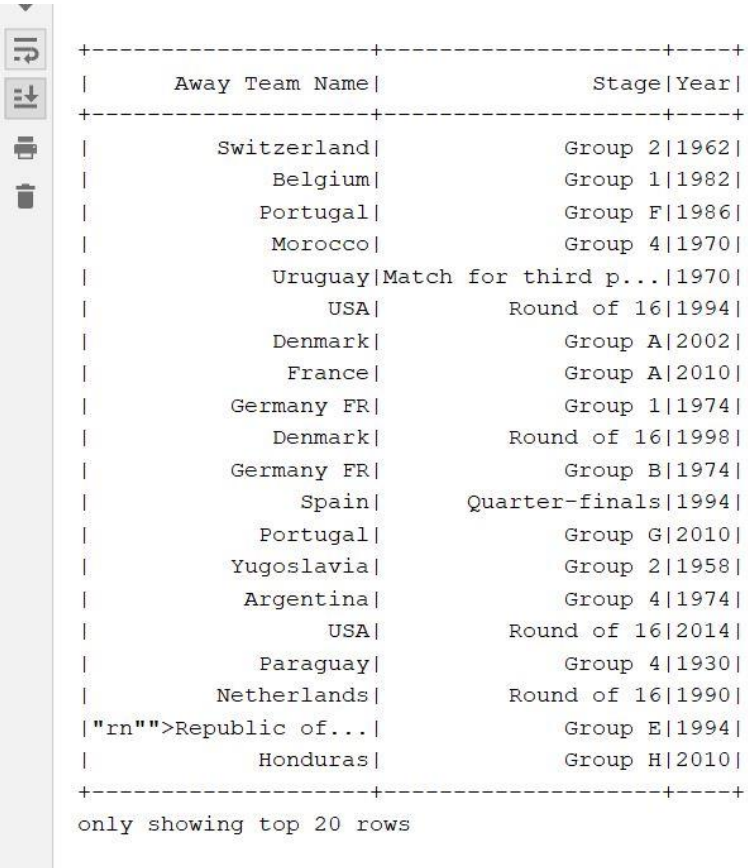
```
+--------------+--------------------+
|count(Stadium)|             Stadium|
+--------------+--------------------+
|             8|          Cuauhtemoc|
|             6|      Parque Central|
|             3|       Idrottsparken|
|             5|         Waldstadion|
|             1|              Friuli|
|             3|       Jose Zorrilla|
|             3|Old Trafford Stadium|
|             3|           San Mames|
|             3|      Miyagi Stadium|
|             6|FIFA World Cup St...|
|             6|Royal Bafokeng Sp...|
|             3|       Nuevo Estadio|
|             4|      Arena Amazonia|
|            11|Nou Camp - Estadi...|
|             4|   Santiago Bernabeu|
|             3| Osaka Nagai Stadium|
|             6|Estadio Jos� Mar�...|
|             2|Ramon Sanchez Piz...|
|             4|     Renato Dall Ara|
|             4|  Pontiac Silverdome|
+--------------+--------------------+
only showing top 20 rows
```



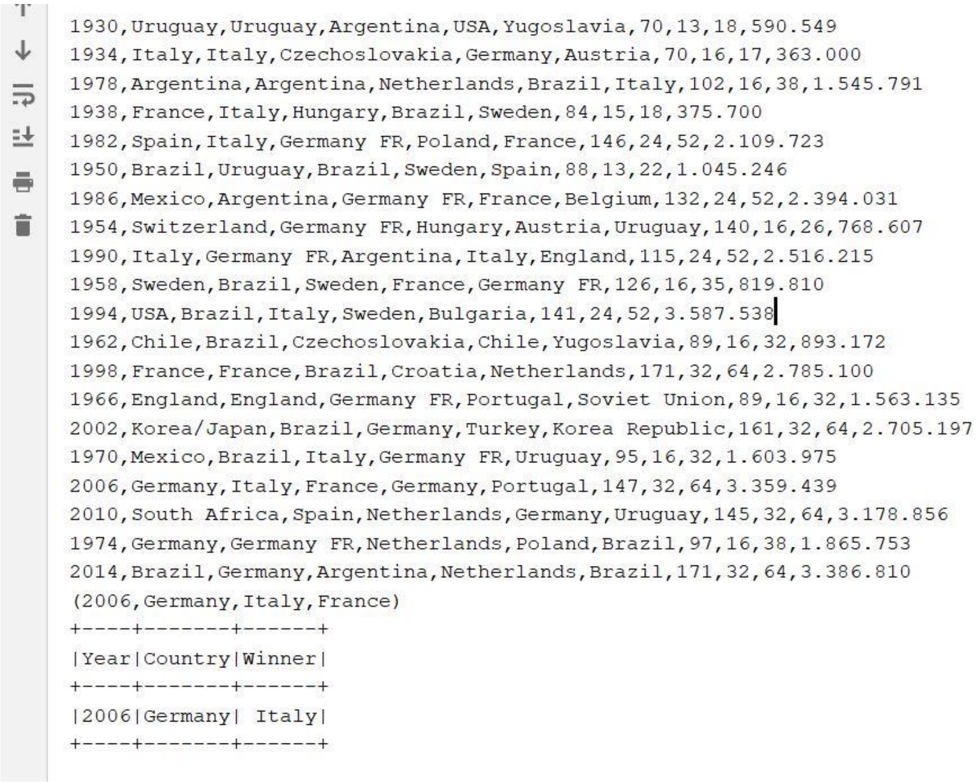
```
+-------------------+--------+
|        Player Name|Position|
+-------------------+--------+
|        Alex THEPOT|      GK|
|    Oscar BONFIGLIO|      GK|
|      Jimmy DOUGLAS|      GK|
|      Arnold BADJOU|      GK|
|     Milovan JAKSIC|      GK|
|               JOEL|      GK|
|     Ion LAPUSNEANU|      GK|
|    Juan VALDIVIESO|      GK|
|       Angel BOSSIO|      GK|
|        Alex THEPOT|      GK|
|     Roberto CORTES|      GK|
|       Isidoro SOTA|      GK|
|     Milovan JAKSIC|      GK|
|     Jesus BERMUDEZ|      GK|
|      Jimmy DOUGLAS|      GK|
|      Modesto DENIS|      GK|
|Enrique BALLESTRERO|      GK|
|       Jorge PARDON|      GK|
|     Roberto CORTES|      GK|
|        Alex THEPOT|      GK|
+-------------------+--------+
only showing top 20 rows
```



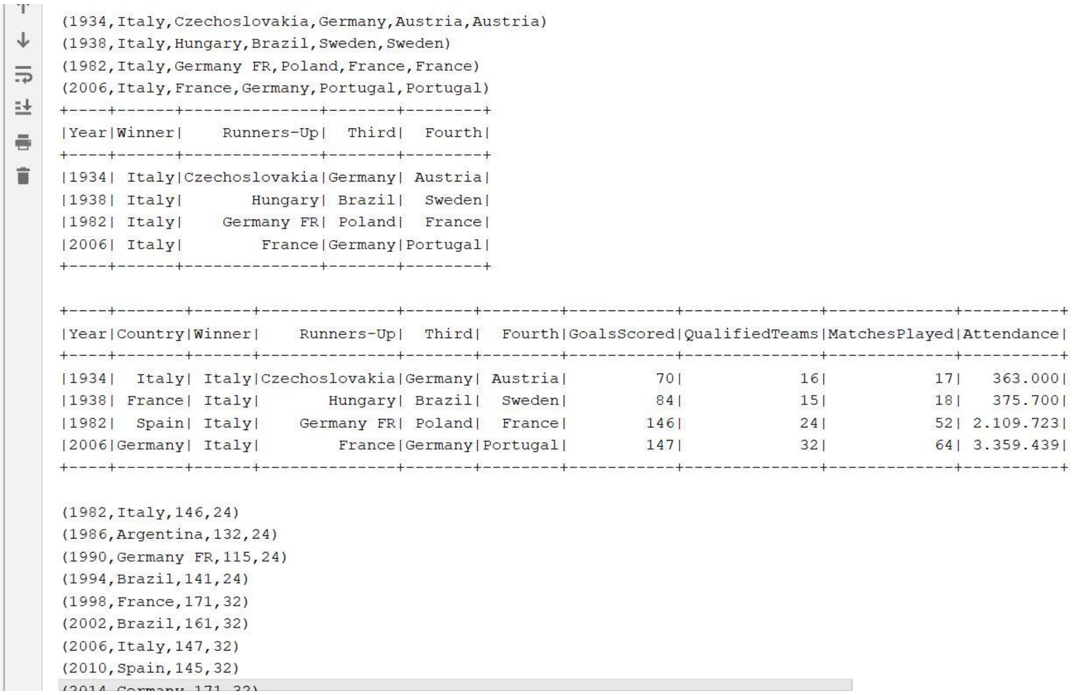
```
+--------------------+--------------------+----+
|      Home Team Name|               Stage|Year|
+--------------------+--------------------+----+
|             Belgium|             Group 1|1982|
|            Portugal|             Group F|1986|
|              Brazil|             Group 3|1962|
|          Germany FR|      Quarter-finals|1966|
|          Germany FR|Match for third p...|1970|
|             Denmark|             Group A|2002|
|              France|             Group A|2010|
|             Hungary|      Quarter-finals|1954|
|          Germany FR|             Group 1|1974|
|             England|             Group F|1990|
|          Germany FR|             Group B|1974|
|           Argentina|         Round of 16|2006|
|            Portugal|             Group G|2010|
|             Germany|         Round of 16|2014|
|          Yugoslavia|             Group 2|1958|
|           Argentina|             Group 4|1974|
|            Paraguay|             Group 4|1930|
|          Germany FR|         Semi-finals|1966|
|"rn"">Republic of...|             Group E|1994|
|            Honduras|             Group H|2010|
+--------------------+--------------------+----+
only showing top 20 rows
```



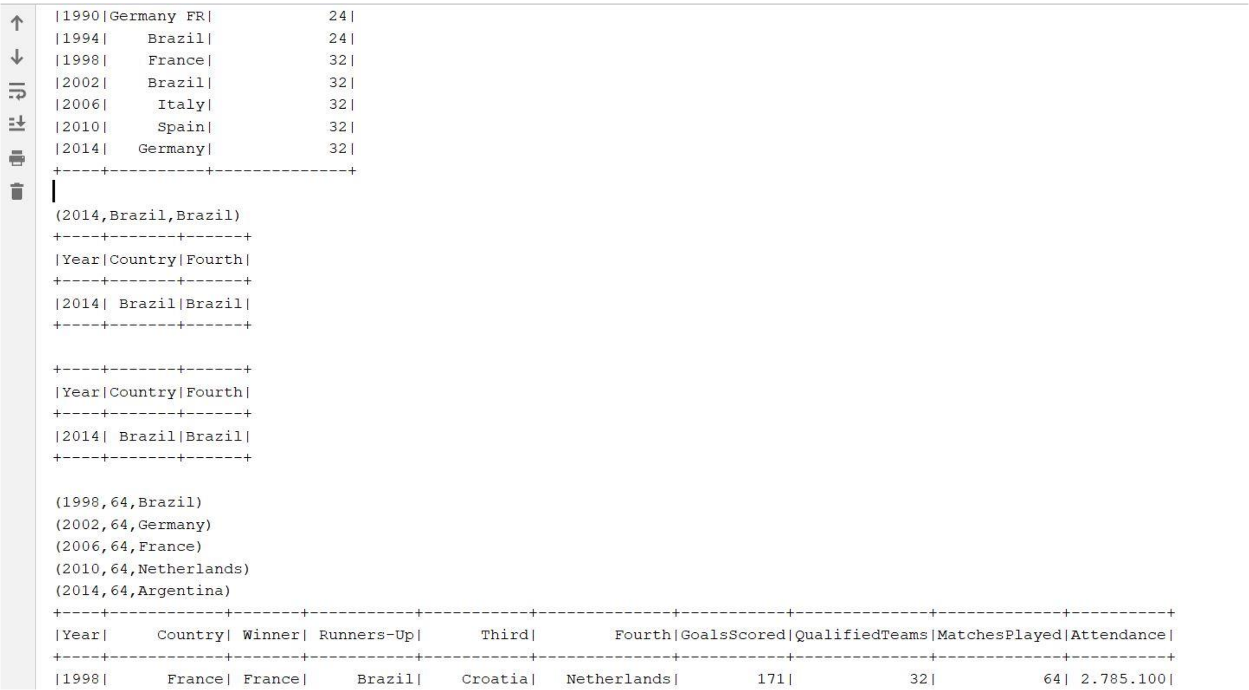
```
+--------------------+--------------------+----+
|      Away Team Name|               Stage|Year|
+--------------------+--------------------+----+
|         Switzerland|             Group 2|1962|
|             Belgium|             Group 1|1982|
|            Portugal|             Group F|1986|
|             Morocco|             Group 4|1970|
|             Uruguay|Match for third p...|1970|
|                 USA|         Round of 16|1994|
|             Denmark|             Group A|2002|
|              France|             Group A|2010|
|          Germany FR|             Group 1|1974|
|             Denmark|         Round of 16|1998|
|          Germany FR|             Group B|1974|
|               Spain|      Quarter-finals|1994|
|            Portugal|             Group G|2010|
|          Yugoslavia|             Group 2|1958|
|           Argentina|             Group 4|1974|
|                 USA|         Round of 16|2014|
|            Paraguay|             Group 4|1930|
|         Netherlands|         Round of 16|1990|
|"rn"">Republic of...|             Group E|1994|
|            Honduras|             Group H|2010|
+--------------------+--------------------+----+
only showing top 20 rows
```



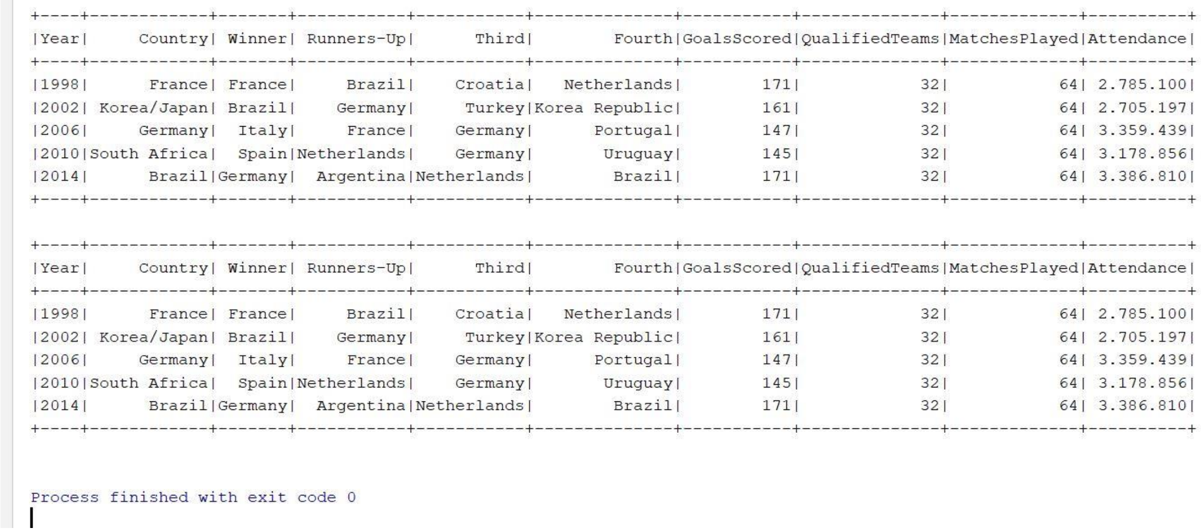
```
1930,Uruguay,Uruguay,Argentina,USA,Yugoslavia,70,13,18,590.549
1934,Italy,Italy,Czechoslovakia,Germany,Austria,70,16,17,363.000
1978,Argentina,Argentina,Netherlands,Brazil,Italy,102,16,38,1.545.791
1938,France,Italy,Hungary,Brazil,Sweden,84,15,18,375.700
1982,Spain,Italy,Germany FR,Poland,France,146,24,52,2.109.723
1950,Brazil,Uruguay,Brazil,Sweden,Spain,88,13,22,1.045.246
1986,Mexico,Argentina,Germany FR,France,Belgium,132,24,52,2.394.031
1954,Switzerland,Germany FR,Hungary,Austria,Uruguay,140,16,26,768.607
1990,Italy,Germany FR,Argentina,Italy,England,115,24,52,2.516.215
1958,Sweden,Brazil,Sweden,France,Germany FR,126,16,35,819.810
1994,USA,Brazil,Italy,Sweden,Bulgaria,141,24,52,3.587.538
1962,Chile,Brazil,Czechoslovakia,Chile,Yugoslavia,89,16,32,893.172
1998,France,France,Brazil,Croatia,Netherlands,171,32,64,2.785.100
1966,England,England,Germany FR,Portugal,Soviet Union,89,16,32,1.563.135
2002,Korea/Japan,Brazil,Germany,Turkey,Korea Republic,161,32,64,2.705.197
1970,Mexico,Brazil,Italy,Germany FR,Uruguay,95,16,32,1.603.975
2006,Germany,Italy,France,Germany,Portugal,147,32,64,3.359.439
2010,South Africa,Spain,Netherlands,Germany,Uruguay,145,32,64,3.178.856
1974,Germany,Germany FR,Netherlands,Poland,Brazil,97,16,38,1.865.753
2014,Brazil,Germany,Argentina,Netherlands,Brazil,171,32,64,3.386.810
(2006,Germany,Italy,France)
+----+-------+------+
|Year|Country|Winner|
+----+-------+------+
|2006|Germany| Italy|
+----+-------+------+
```



```
(1934,Italy,Czechoslovakia,Germany,Austria,Austria)
(1938,Italy,Hungary,Brazil,Sweden,Sweden)
(1982,Italy,Germany FR,Poland,France,France)
(2006,Italy,France,Germany,Portugal,Portugal)
+----+------+--------------+-------+--------+
|Year|Winner|    Runners-Up|  Third|  Fourth|
+----+------+--------------+-------+--------+
|1934| Italy|Czechoslovakia|Germany| Austria|
|1938| Italy|       Hungary| Brazil|  Sweden|
|1982| Italy|    Germany FR| Poland|  France|
|2006| Italy|        France|Germany|Portugal|
+----+------+--------------+-------+--------+

+----+-------+------+--------------+-------+--------+-----------+--------------+-------------+----------+
|Year|Country|Winner|    Runners-Up|  Third|  Fourth|GoalsScored|QualifiedTeams|MatchesPlayed|Attendance|
+----+-------+------+--------------+-------+--------+-----------+--------------+-------------+----------+
|1934|  Italy| Italy|Czechoslovakia|Germany| Austria|         70|            16|           17|   363.000|
|1938| France| Italy|       Hungary| Brazil|  Sweden|         84|            15|           18|   375.700|
|1982|  Spain| Italy|    Germany FR| Poland|  France|        146|            24|           52| 2.109.723|
|2006|Germany| Italy|        France|Germany|Portugal|        147|            32|           64| 3.359.439|
+----+-------+------+--------------+-------+--------+-----------+--------------+-------------+----------+

(1982,Italy,146,24)
(1986,Argentina,132,24)
(1990,Germany FR,115,24)
(1994,Brazil,141,24)
(1998,France,171,32)
(2002,Brazil,161,32)
(2006,Italy,147,32)
(2010,Spain,145,32)
(2014,Germany,171,32)
```



```
|1990|Germany FR|            24|
|1994|    Brazil|            24|
|1998|    France|            32|
|2002|    Brazil|            32|
|2006|     Italy|            32|
|2010|     Spain|            32|
|2014|   Germany|            32|
+----+----------+--------------+

(2014,Brazil,Brazil)
+----+-------+------+
|Year|Country|Fourth|
+----+-------+------+
|2014| Brazil|Brazil|
+----+-------+------+

+----+-------+------+
|Year|Country|Fourth|
+----+-------+------+
|2014| Brazil|Brazil|
+----+-------+------+

(1998,64,Brazil)
(2002,64,Germany)
(2006,64,France)
(2010,64,Netherlands)
(2014,64,Argentina)
+----+------------+-------+-----------+-----------+--------------+-----------+--------------+-------------+----------+
|Year|     Country| Winner| Runners-Up|      Third|        Fourth|GoalsScored|QualifiedTeams|MatchesPlayed|Attendance|
+----+------------+-------+-----------+-----------+--------------+-----------+--------------+-------------+----------+
|1998|      France| France|     Brazil|    Croatia|   Netherlands|        171|            32|           64| 2.785.100|
```

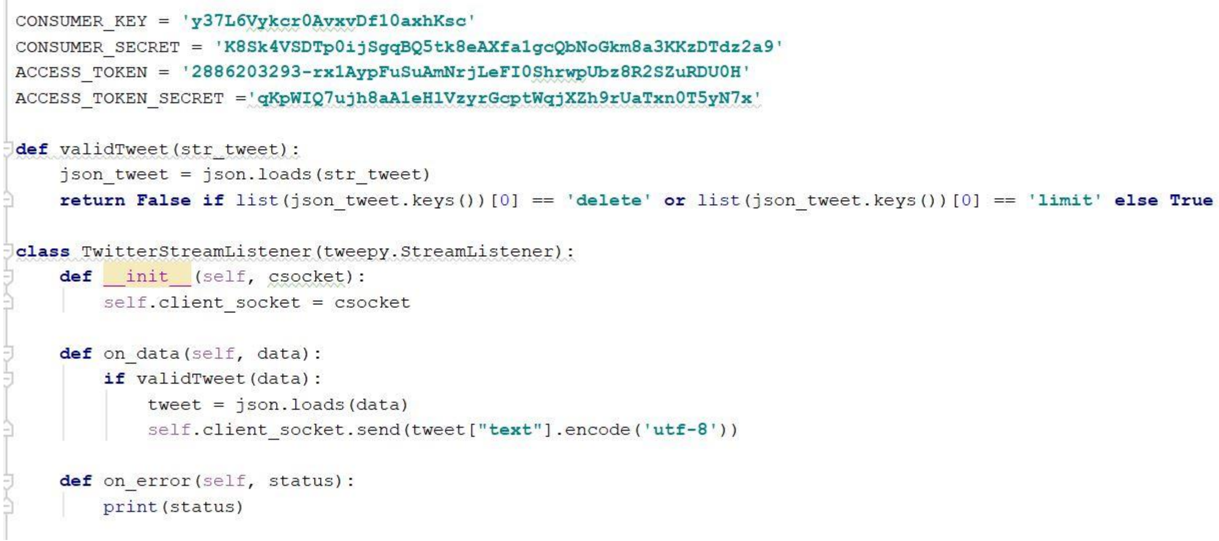


```
+----+------------+-------+-----------+-----------+--------------+-----------+--------------+-------------+----------+
|Year|     Country| Winner| Runners-Up|      Third|        Fourth|GoalsScored|QualifiedTeams|MatchesPlayed|Attendance|
+----+------------+-------+-----------+-----------+--------------+-----------+--------------+-------------+----------+
|1998|      France| France|     Brazil|    Croatia|   Netherlands|        171|            32|           64| 2.785.100|
|2002| Korea/Japan| Brazil|    Germany|     Turkey|Korea Republic|        161|            32|           64| 2.705.197|
|2006|     Germany|  Italy|     France|    Germany|      Portugal|        147|            32|           64| 3.359.439|
|2010|South Africa|  Spain|Netherlands|    Germany|       Uruguay|        145|            32|           64| 3.178.856|
|2014|      Brazil|Germany|  Argentina|Netherlands|        Brazil|        171|            32|           64| 3.386.810|
+----+------------+-------+-----------+-----------+--------------+-----------+--------------+-------------+----------+

+----+------------+-------+-----------+-----------+--------------+-----------+--------------+-------------+----------+
|Year|     Country| Winner| Runners-Up|      Third|        Fourth|GoalsScored|QualifiedTeams|MatchesPlayed|Attendance|
+----+------------+-------+-----------+-----------+--------------+-----------+--------------+-------------+----------+
|1998|      France| France|     Brazil|    Croatia|   Netherlands|        171|            32|           64| 2.785.100|
|2002| Korea/Japan| Brazil|    Germany|     Turkey|Korea Republic|        161|            32|           64| 2.705.197|
|2006|     Germany|  Italy|     France|    Germany|      Portugal|        147|            32|           64| 3.359.439|
|2010|South Africa|  Spain|Netherlands|    Germany|       Uruguay|        145|            32|           64| 3.178.856|
|2014|      Brazil|Germany|  Argentina|Netherlands|        Brazil|        171|            32|           64| 3.386.810|
+----+------------+-------+-----------+-----------+--------------+-----------+--------------+-------------+----------+


Process finished with exit code 0
```

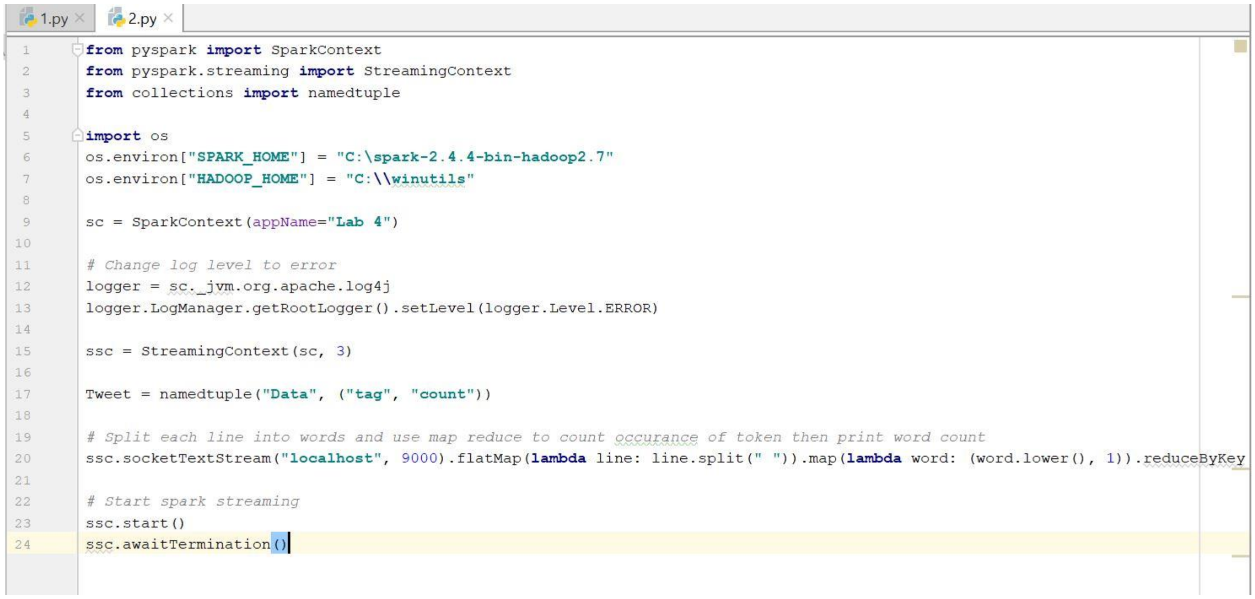
Part 3

Aim :

Spark Streaming : Perform Word-Count on Twitter Streaming Data using Spark.

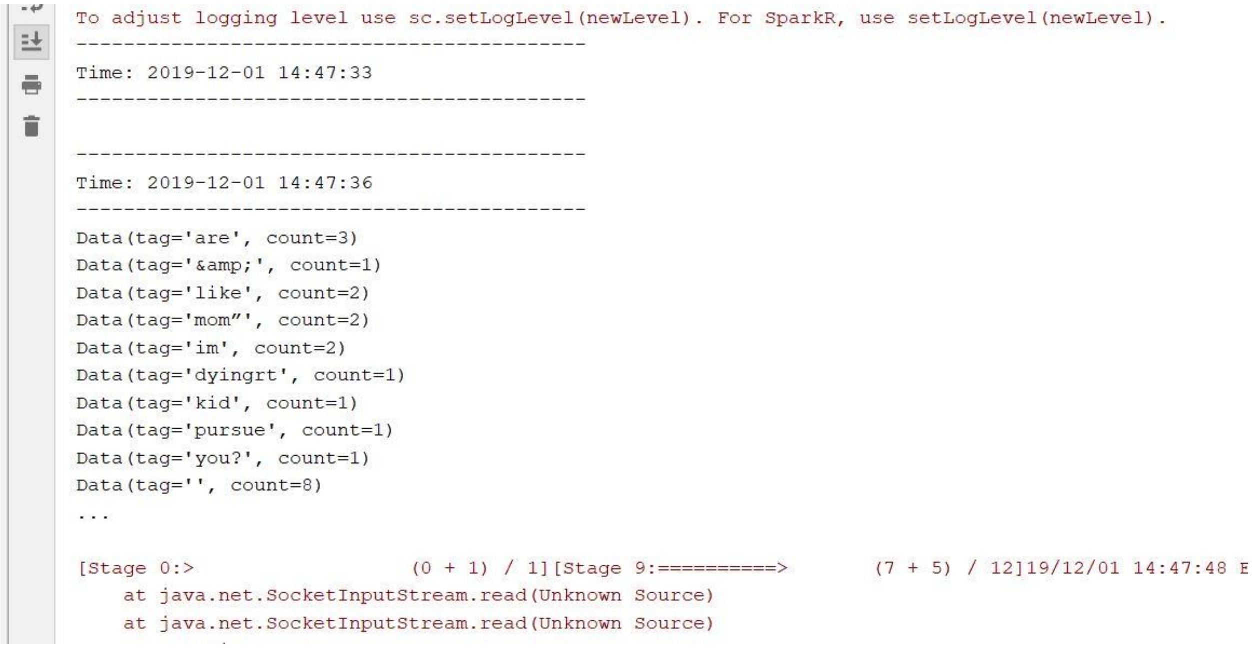


```python
CONSUMER_KEY = 'y37L6Vykcr0AvxvDf10axhKsc'
CONSUMER_SECRET = 'K8Sk4VSDTp0ijSgqBQ5tk8eAXfa1gcQbNoGkm8a3KKzDTdz2a9'
ACCESS_TOKEN = '2886203293-rx1AypFuSuAmNrjLeFI0ShrwpUbz8R2SZuRDU0H'
ACCESS_TOKEN_SECRET ='qKpWIQ7ujh8aA1eH1VzyrGcptWqjXZh9rUaTxn0T5yN7x'

def validTweet(str_tweet):
    json_tweet = json.loads(str_tweet)
    return False if list(json_tweet.keys())[0] == 'delete' or list(json_tweet.keys())[0] == 'limit' else True

class TwitterStreamListener(tweepy.StreamListener):
    def __init__(self, csocket):
        self.client_socket = csocket

    def on_data(self, data):
        if validTweet(data):
            tweet = json.loads(data)
            self.client_socket.send(tweet["text"].encode('utf-8'))

    def on_error(self, status):
        print(status)
```



```python
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from collections import namedtuple

import os
os.environ["SPARK_HOME"] = "C:\spark-2.4.4-bin-hadoop2.7"
os.environ["HADOOP_HOME"] = "C:\\winutils"

sc = SparkContext(appName="Lab 4")

# Change log level to error
logger = sc._jvm.org.apache.log4j
logger.LogManager.getRootLogger().setLevel(logger.Level.ERROR)

ssc = StreamingContext(sc, 3)

Tweet = namedtuple("Data", ("tag", "count"))

# Split each line into words and use map reduce to count occurance of token then print word count
ssc.socketTextStream("localhost", 9000).flatMap(lambda line: line.split(" ")).map(lambda word: (word.lower(), 1)).reduceByKey

# Start spark streaming
ssc.start()
ssc.awaitTermination()
```

Output :



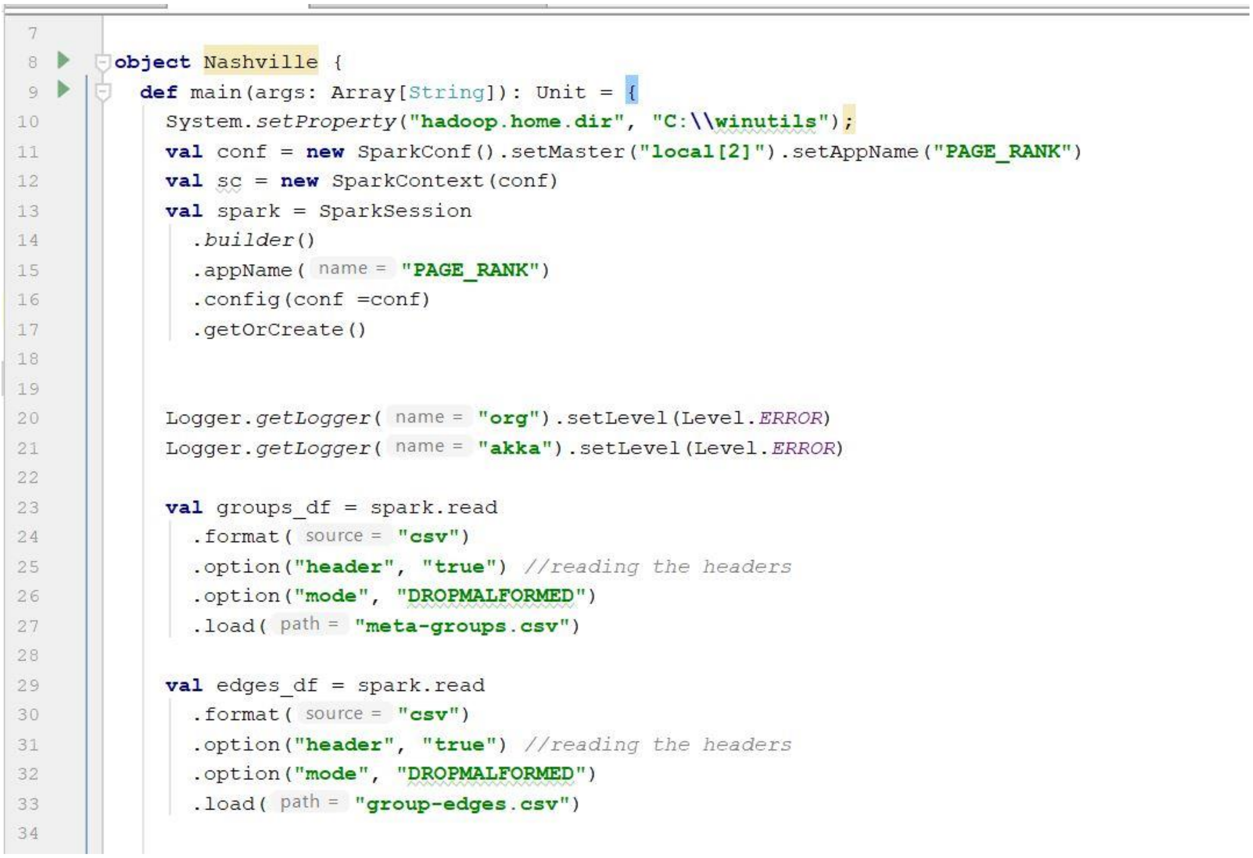
```
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
-------------------------------------------
Time: 2019-12-01 14:47:33
-------------------------------------------

-------------------------------------------
Time: 2019-12-01 14:47:36
-------------------------------------------
Data(tag='are', count=3)
Data(tag='&amp;', count=1)
Data(tag='like', count=2)
Data(tag='mom”', count=2)
Data(tag='im', count=2)
Data(tag='dyingrt', count=1)
Data(tag='kid', count=1)
Data(tag='pursue', count=1)
Data(tag='you?', count=1)
Data(tag='', count=8)
...

[Stage 0:>                  (0 + 1) / 1][Stage 9:==========>       (7 + 5) / 12]19/12/01 14:47:48 E
    at java.net.SocketInputStream.read(Unknown Source)
    at java.net.SocketInputStream.read(Unknown Source)
```

Part 4

Aim :

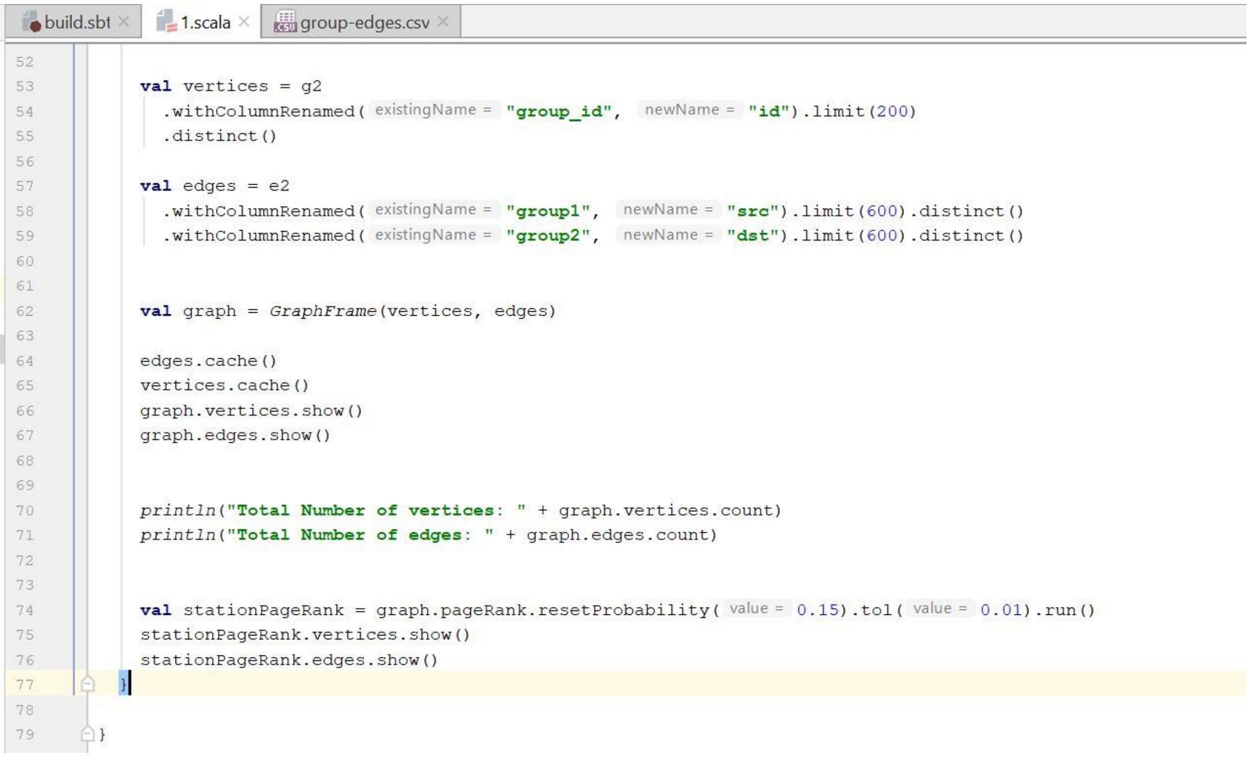
Spark Graphx Task

We used the dataset Nashville for this task.



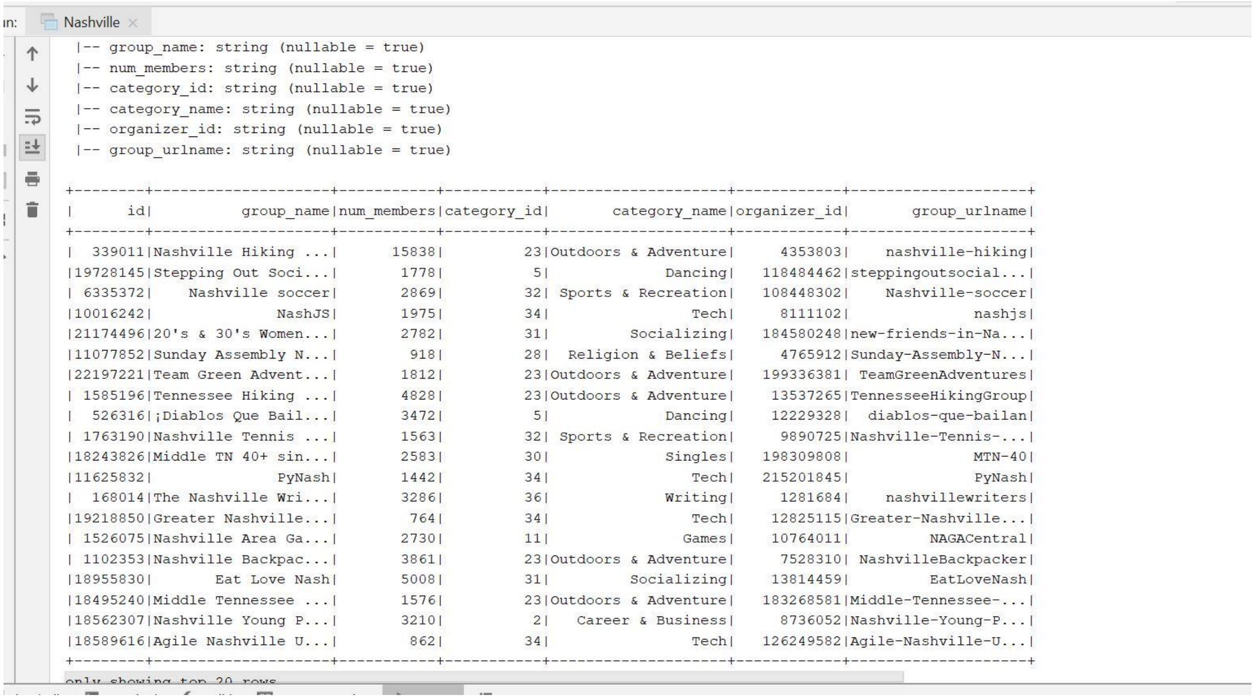
```scala
object Nashville {
  def main(args: Array[String]): Unit = {
    System.setProperty("hadoop.home.dir", "C:\\winutils");
    val conf = new SparkConf().setMaster("local[2]").setAppName("PAGE_RANK")
    val sc = new SparkContext(conf)
    val spark = SparkSession
      .builder()
      .appName( name = "PAGE_RANK")
      .config(conf =conf)
      .getOrCreate()


    Logger.getLogger( name = "org").setLevel(Level.ERROR)
    Logger.getLogger( name = "akka").setLevel(Level.ERROR)

    val groups_df = spark.read
      .format( source = "csv")
      .option("header", "true") //reading the headers
      .option("mode", "DROPMALFORMED")
      .load( path = "meta-groups.csv")

    val edges_df = spark.read
      .format( source = "csv")
      .option("header", "true") //reading the headers
      .option("mode", "DROPMALFORMED")
      .load( path = "group-edges.csv")

```

1. Perform Page Rank

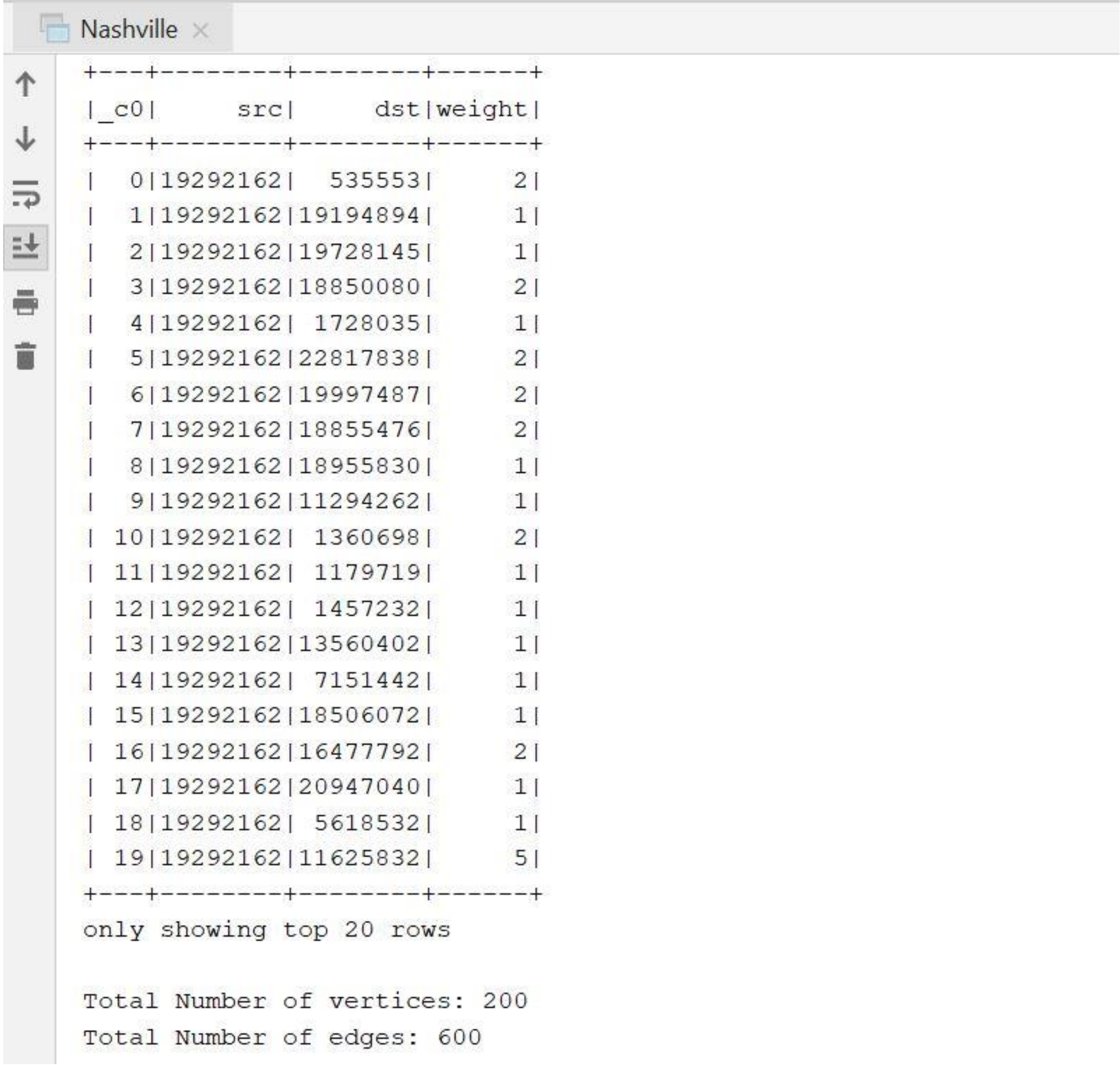


```scala

    val vertices = g2
      .withColumnRenamed( existingName = "group_id",  newName = "id").limit(200)
      .distinct()

    val edges = e2
      .withColumnRenamed( existingName = "group1",  newName = "src").limit(600).distinct()
      .withColumnRenamed( existingName = "group2",  newName = "dst").limit(600).distinct()


    val graph = GraphFrame(vertices, edges)

    edges.cache()
    vertices.cache()
    graph.vertices.show()
    graph.edges.show()


    println("Total Number of vertices: " + graph.vertices.count)
    println("Total Number of edges: " + graph.edges.count)


    val stationPageRank = graph.pageRank.resetProbability( value = 0.15).tol( value = 0.01).run()
    stationPageRank.vertices.show()
    stationPageRank.edges.show()
  }

}
```

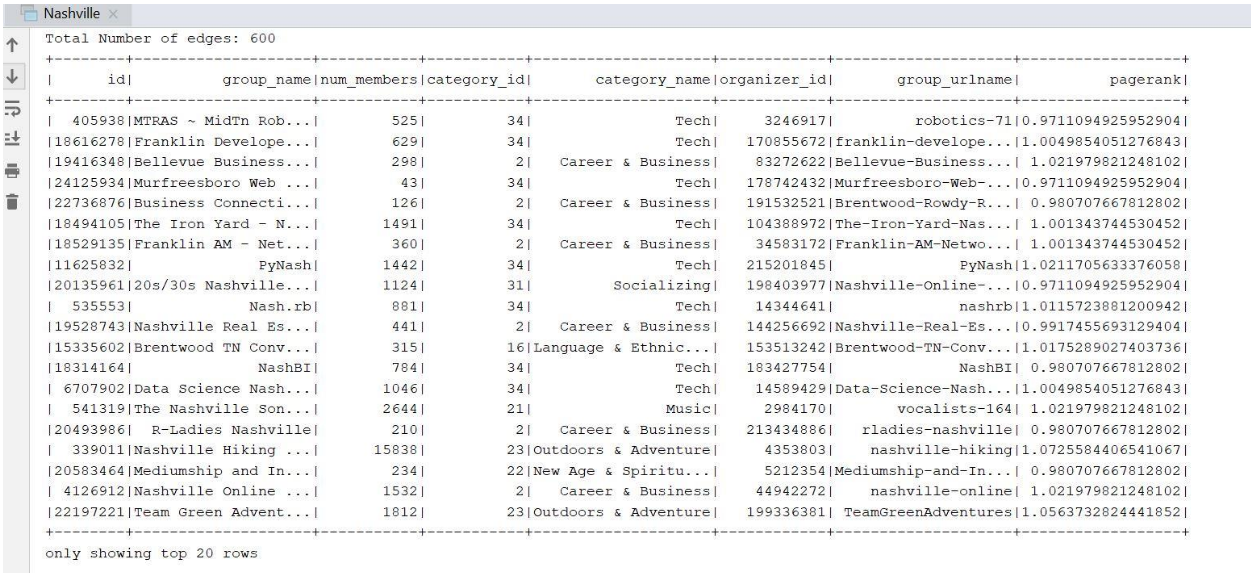
Output :



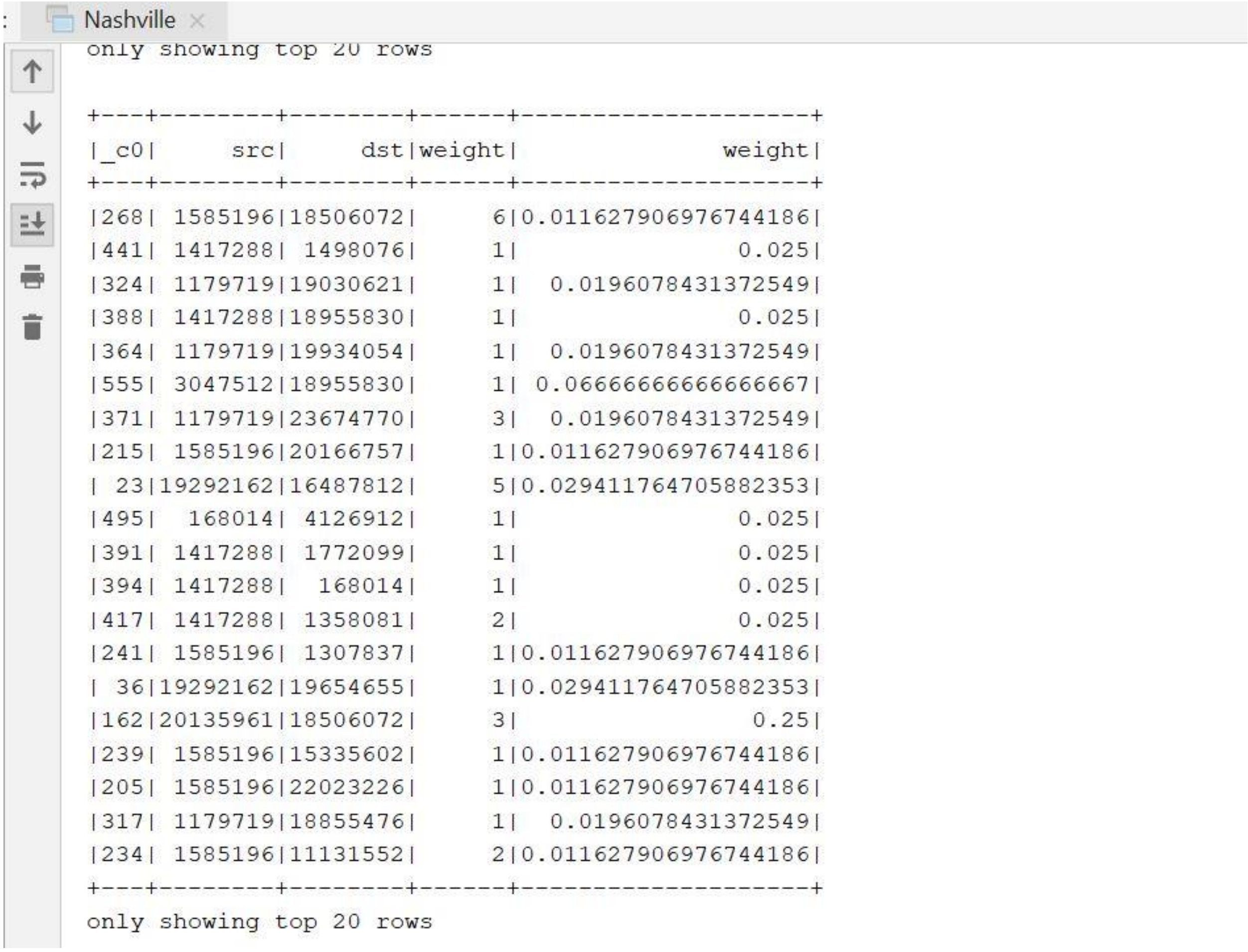
```
 |-- group_name: string (nullable = true)
 |-- num_members: string (nullable = true)
 |-- category_id: string (nullable = true)
 |-- category_name: string (nullable = true)
 |-- organizer_id: string (nullable = true)
 |-- group_urlname: string (nullable = true)

+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+
|      id|          group_name|num_members|category_id|       category_name|organizer_id|       group_urlname|
+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+
|  339011|Nashville Hiking ...|      15838|         23|Outdoors & Adventure|     4353803|    nashville-hiking|
|19728145|Stepping Out Soci...|       1778|          5|             Dancing|   118484462|steppingoutsocial...|
| 6335372|    Nashville soccer|       2869|         32| Sports & Recreation|   108448302|    Nashville-soccer|
|10016242|              NashJS|       1975|         34|                Tech|     8111102|              nashjs|
|21174496|20's & 30's Women...|       2782|         31|         Socializing|   184580248|new-friends-in-Na...|
|11077852|Sunday Assembly N...|        918|         28|  Religion & Beliefs|     4765912|Sunday-Assembly-N...|
|22197221|Team Green Advent...|       1812|         23|Outdoors & Adventure|   199336381| TeamGreenAdventures|
| 1585196|Tennessee Hiking ...|       4828|         23|Outdoors & Adventure|    13537265|TennesseeHikingGroup|
|  526316|¡Diablos Que Bail...|       3472|          5|             Dancing|    12229328|  diablos-que-bailan|
| 1763190|Nashville Tennis ...|       1563|         32| Sports & Recreation|     9890725|Nashville-Tennis-...|
|18243826|Middle TN 40+ sin...|       2583|         30|             Singles|   198309808|              MTN-40|
|11625832|              PyNash|       1442|         34|                Tech|   215201845|              PyNash|
|  168014|The Nashville Wri...|       3286|         36|             Writing|     1281684|    nashvillewriters|
|19218850|Greater Nashville...|        764|         34|                Tech|    12825115|Greater-Nashville...|
| 1526075|Nashville Area Ga...|       2730|         11|               Games|    10764011|         NAGACentral|
| 1102353|Nashville Backpac...|       3861|         23|Outdoors & Adventure|     7528310| NashvilleBackpacker|
|18955830|       Eat Love Nash|       5008|         31|         Socializing|    13814459|         EatLoveNash|
|18495240|Middle Tennessee ...|       1576|         23|Outdoors & Adventure|   183268581|Middle-Tennessee-...|
|18562307|Nashville Young P...|       3210|          2|   Career & Business|     8736052|Nashville-Young-P...|
|18589616|Agile Nashville U...|        862|         34|                Tech|   126249582|Agile-Nashville-U...|
+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+
only showing top 20 rows
```



```
+---+--------+--------+------+
|_c0|     src|     dst|weight|
+---+--------+--------+------+
|  0|19292162|  535553|     2|
|  1|19292162|19194894|     1|
|  2|19292162|19728145|     1|
|  3|19292162|18850080|     2|
|  4|19292162| 1728035|     1|
|  5|19292162|22817838|     2|
|  6|19292162|19997487|     2|
|  7|19292162|18855476|     2|
|  8|19292162|18955830|     1|
|  9|19292162|11294262|     1|
| 10|19292162| 1360698|     2|
| 11|19292162| 1179719|     1|
| 12|19292162| 1457232|     1|
| 13|19292162|13560402|     1|
| 14|19292162| 7151442|     1|
| 15|19292162|18506072|     1|
| 16|19292162|16477792|     2|
| 17|19292162|20947040|     1|
| 18|19292162| 5618532|     1|
| 19|19292162|11625832|     5|
+---+--------+--------+------+
only showing top 20 rows

Total Number of vertices: 200
Total Number of edges: 600
```



```
Total Number of edges: 600
+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+------------------+
|      id|          group_name|num_members|category_id|       category_name|organizer_id|       group_urlname|          pagerank|
+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+------------------+
|  405938|MTRAS ~ MidTn Rob...|        525|         34|                Tech|     3246917|         robotics-71|0.9711094925952904|
|18616278|Franklin Develope...|        629|         34|                Tech|   170855672|franklin-develope...|1.0049854051276843|
|19416348|Bellevue Business...|        298|          2|   Career & Business|    83272622|Bellevue-Business...| 1.021979821248102|
|24125934|Murfreesboro Web ...|         43|         34|                Tech|   178742432|Murfreesboro-Web-...|0.9711094925952904|
|22736876|Business Connecti...|        126|          2|   Career & Business|   191532521|Brentwood-Rowdy-R...| 0.980707667812802|
|18494105|The Iron Yard - N...|       1491|         34|                Tech|   104388972|The-Iron-Yard-Nas...| 1.001343744530452|
|18529135|Franklin AM - Net...|        360|          2|   Career & Business|    34583172|Franklin-AM-Netwo...| 1.001343744530452|
|11625832|              PyNash|       1442|         34|                Tech|   215201845|              PyNash|1.0211705633376058|
|20135961|20s/30s Nashville...|       1124|         31|         Socializing|   198403977|Nashville-Online-...|0.9711094925952904|
|  535553|             Nash.rb|        881|         34|                Tech|    14344641|              nashrb|1.0115723881200942|
|19528743|Nashville Real Es...|        441|          2|   Career & Business|   144256692|Nashville-Real-Es...|0.9917455693129404|
|15335602|Brentwood TN Conv...|        315|         16|Language & Ethnic...|   153513242|Brentwood-TN-Conv...|1.0175289027403736|
|18314164|              NashBI|        784|         34|                Tech|   183427754|              NashBI| 0.980707667812802|
| 6707902|Data Science Nash...|       1046|         34|                Tech|    14589429|Data-Science-Nash...|1.0049854051276843|
|  541319|The Nashville Son...|       2644|         21|               Music|     2984170|       vocalists-164| 1.021979821248102|
|20493986|  R-Ladies Nashville|        210|          2|   Career & Business|   213434886|   rladies-nashville| 0.980707667812802|
|  339011|Nashville Hiking ...|      15838|         23|Outdoors & Adventure|     4353803|    nashville-hiking|1.0725584406541067|
|20583464|Mediumship and In...|        234|         22|New Age & Spiritu...|     5212354|Mediumship-and-In...| 0.980707667812802|
| 4126912|Nashville Online ...|       1532|          2|   Career & Business|    44942272|    nashville-online| 1.021979821248102|
|22197221|Team Green Advent...|       1812|         23|Outdoors & Adventure|   199336381| TeamGreenAdventures|1.0563732824441852|
+--------+--------------------+-----------+-----------+--------------------+------------+--------------------+------------------+
only showing top 20 rows
```



```
only showing top 20 rows

+---+--------+--------+------+--------------------+
|_c0|     src|     dst|weight|              weight|
+---+--------+--------+------+--------------------+
|268| 1585196|18506072|     6|0.011627906976744186|
|441| 1417288| 1498076|     1|               0.025|
|324| 1179719|19030621|     1|  0.0196078431372549|
|388| 1417288|18955830|     1|               0.025|
|364| 1179719|19934054|     1|  0.0196078431372549|
|555| 3047512|18955830|     1| 0.06666666666666667|
|371| 1179719|23674770|     3|  0.0196078431372549|
|215| 1585196|20166757|     1|0.011627906976744186|
| 23|19292162|16487812|     5|0.029411764705882353|
|495|  168014| 4126912|     1|               0.025|
|391| 1417288| 1772099|     1|               0.025|
|394| 1417288|  168014|     1|               0.025|
|417| 1417288| 1358081|     2|               0.025|
|241| 1585196| 1307837|     1|0.011627906976744186|
| 36|19292162|19654655|     1|0.029411764705882353|
|162|20135961|18506072|     3|                0.25|
|239| 1585196|15335602|     1|0.011627906976744186|
|205| 1585196|22023226|     1|0.011627906976744186|
|317| 1179719|18855476|     1|  0.0196078431372549|
|234| 1585196|11131552|     2|0.011627906976744186|
+---+--------+--------+------+--------------------+
only showing top 20 rows
```

2. State importance of using graphx on the chosen dataset.

Graphx are mainly used for distributed processing of graphs. For example, where a graph is very large with huge no:of vertices and edges then it is difficult to process on a single state machine. Then we need to use parallel computation. Here, we used group-id to produce vertices and group-1,group-2 taken from the dataset which is used to produce edges for the graphs.