ASSIGNMENT 28.1

Aviation data analysis

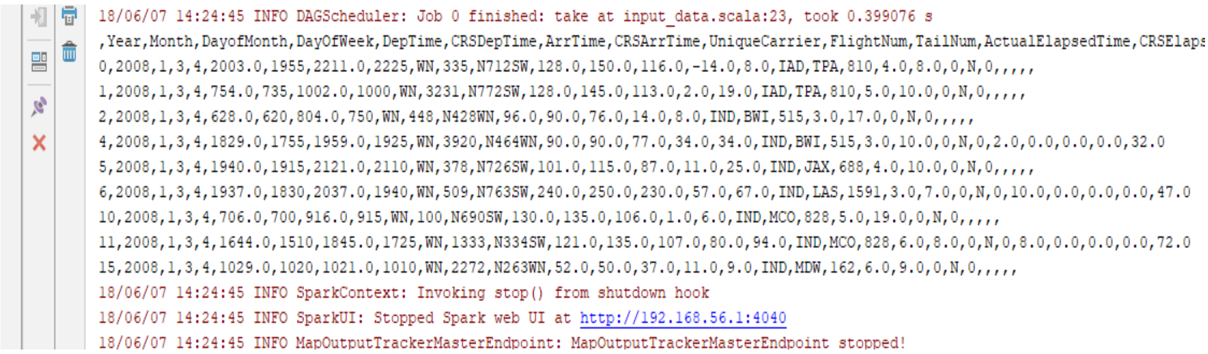
The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, cancelled, and diverted flights appears in DOT's monthly Air Travel Consumer Report, published about 30 days after the month's end, as well as in summary tables posted on this website. Summary statistics and raw data are made available to the public at the time the Air Travel Consumer Report is released.

To solve the below three problem I created the three scala file with object name **Top_5_Destinations, Most_Cancellations** and **Maximum_Diversion**

Description of all codes are explained in the code file.

Below screenshots, show the input dataset and output obtained by the code for each problem in IntelliJIDEA application

Below screenshot shows the sample data from input file **DelyedFlight.csv**

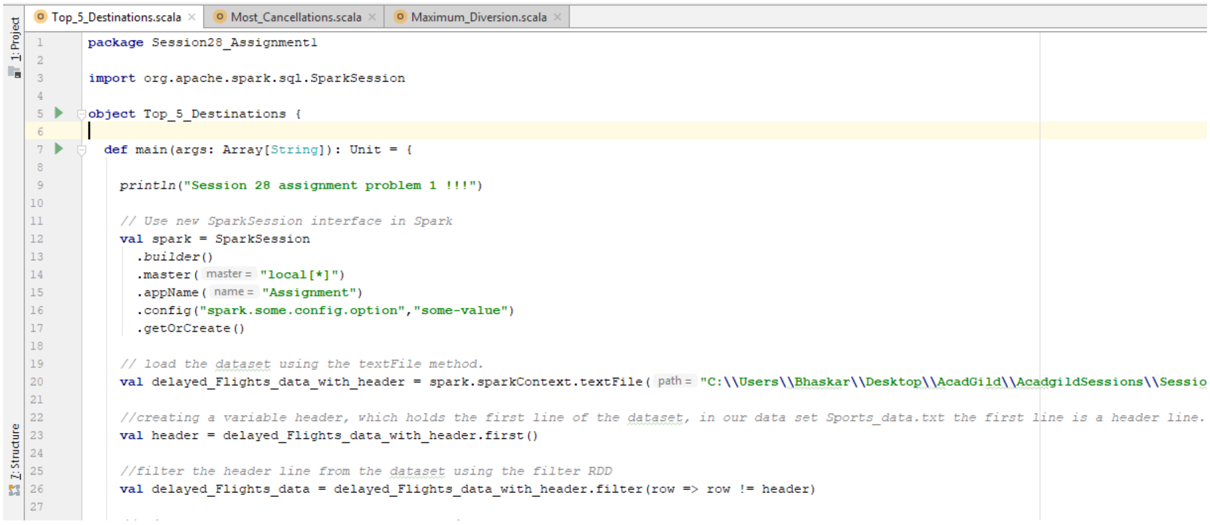


```
18/06/07 14:24:45 INFO DAGScheduler: Job 0 finished: take at input_data.scala:23, took 0.399076 s
,Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElaps
0,2008,1,3,4,2003.0,1955,2211.0,2225,WN,335,N712SW,128.0,150.0,116.0,-14.0,8.0,IAD,TPA,810,4.0,8.0,0,N,0,,,,,
1,2008,1,3,4,754.0,735,1002.0,1000,WN,3231,N772SW,128.0,145.0,113.0,2.0,19.0,IAD,TPA,810,5.0,10.0,0,N,0,,,,,
2,2008,1,3,4,628.0,620,804.0,750,WN,448,N428WN,96.0,90.0,76.0,14.0,8.0,IND,BWI,515,3.0,17.0,0,N,0,,,,,
4,2008,1,3,4,1829.0,1755,1959.0,1925,WN,3920,N464WN,90.0,90.0,77.0,34.0,34.0,IND,BWI,515,3.0,10.0,0,N,0,2.0,0.0,0.0,0.0,32.0
5,2008,1,3,4,1940.0,1915,2121.0,2110,WN,378,N726SW,101.0,115.0,87.0,11.0,25.0,IND,JAX,688,4.0,10.0,0,N,0,,,,,
6,2008,1,3,4,1937.0,1830,2037.0,1940,WN,509,N763SW,240.0,250.0,230.0,57.0,67.0,IND,LAS,1591,3.0,7.0,0,N,0,10.0,0.0,0.0,0.0,47.0
10,2008,1,3,4,706.0,700,916.0,915,WN,100,N690SW,130.0,135.0,106.0,1.0,6.0,IND,MCO,828,5.0,19.0,0,N,0,,,,,
11,2008,1,3,4,1644.0,1510,1845.0,1725,WN,1333,N334SW,121.0,135.0,107.0,80.0,94.0,IND,MCO,828,6.0,8.0,0,N,0,8.0,0.0,0.0,0.0,72.0
15,2008,1,3,4,1029.0,1020,1021.0,1010,WN,2272,N263WN,52.0,50.0,37.0,11.0,9.0,IND,MDW,162,6.0,9.0,0,N,0,,,,,
18/06/07 14:24:45 INFO SparkContext: Invoking stop() from shutdown hook
18/06/07 14:24:45 INFO SparkUI: Stopped Spark web UI at http://192.168.56.1:4040
18/06/07 14:24:45 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
```

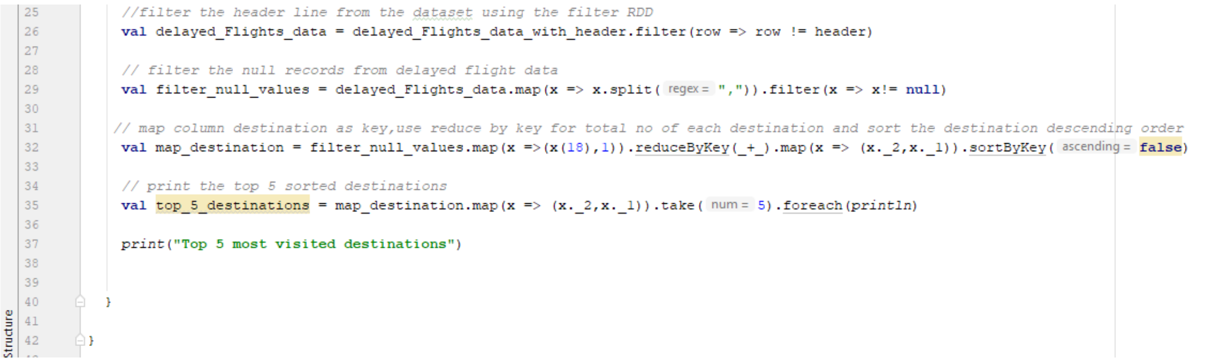
Problem Statement 1

Find out the top 5 most visited destinations.

Below screenshot shows the required code for above problem

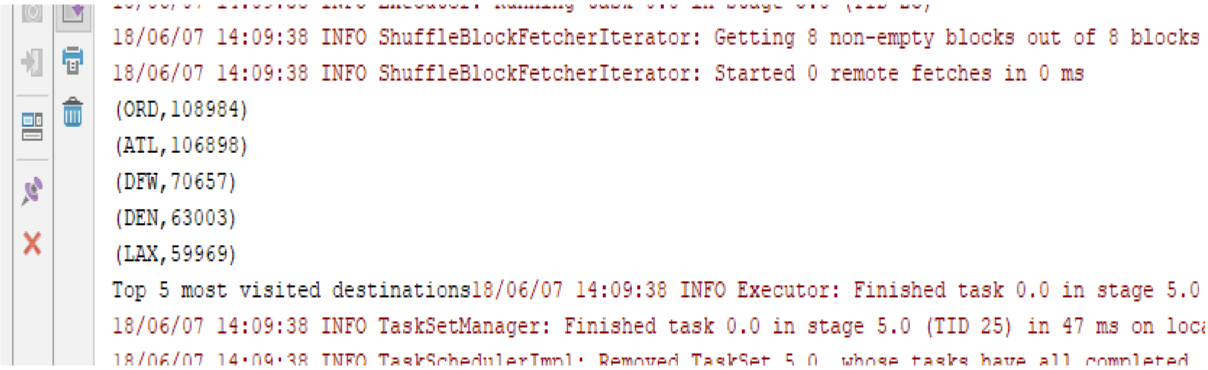


```
package Session28_Assignment1

import org.apache.spark.sql.SparkSession

object Top_5_Destinations {

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

    println("Session 28 assignment problem 1 !!!")

    // Use new SparkSession interface in Spark
    val spark = SparkSession
      .builder()
      .master( master = "local[*]")
      .appName( name = "Assignment")
      .config("spark.some.config.option","some-value")
      .getOrCreate()

    // load the dataset using the textFile method.
    val delayed_Flights_data_with_header = spark.sparkContext.textFile( path = "C:\\Users\\Bhaskar\\Desktop\\AcadGild\\AcadgildSessions\\Sessio

    //creating a variable header, which holds the first line of the dataset, in our data set Sports_data.txt the first line is a header line.
    val header = delayed_Flights_data_with_header.first()

    //filter the header line from the dataset using the filter RDD
    val delayed_Flights_data = delayed_Flights_data_with_header.filter(row => row != header)

```



```
    //filter the header line from the dataset using the filter RDD
    val delayed_Flights_data = delayed_Flights_data_with_header.filter(row => row != header)

    // filter the null records from delayed flight data
    val filter_null_values = delayed_Flights_data.map(x => x.split( regex = ",")).filter(x => x!= null)

   // map column destination as key,use reduce by key for total no of each destination and sort the destination descending order
    val map_destination = filter_null_values.map(x =>(x(18),1)).reduceByKey(_+_).map(x => (x._2,x._1)).sortByKey( ascending = false)

    // print the top 5 sorted destinations
    val top_5_destinations = map_destination.map(x => (x._2,x._1)).take( num = 5).foreach(println)

    print("Top 5 most visited destinations")


  }

}
```

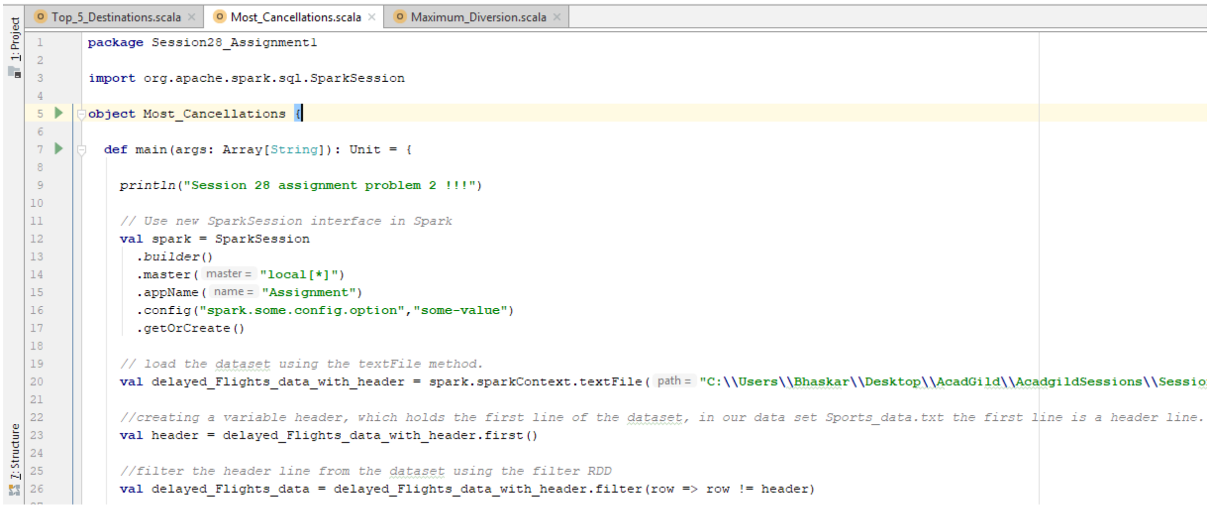
Below screenshot shows the output:



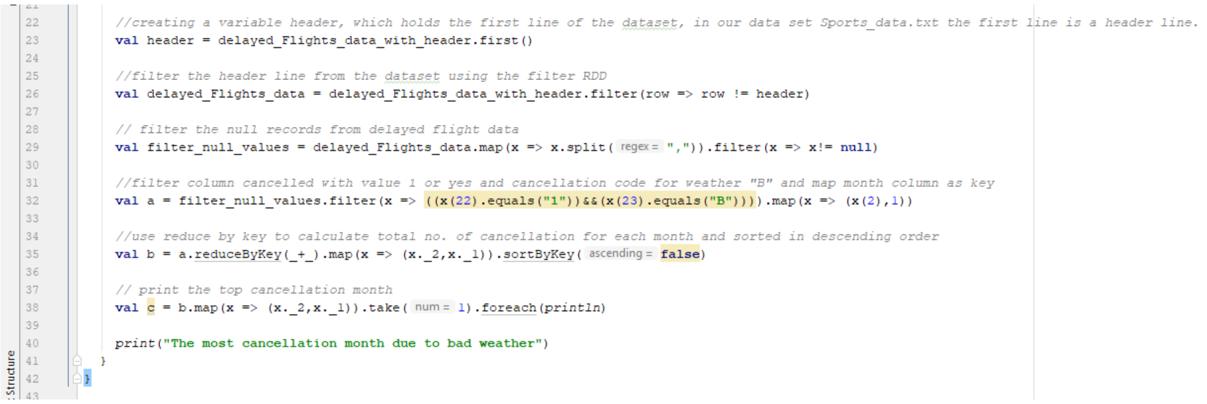
```
18/06/07 14:09:38 INFO Executor: Running task 0.0 in stage 5.0 (TID 25)
18/06/07 14:09:38 INFO ShuffleBlockFetcherIterator: Getting 8 non-empty blocks out of 8 blocks
18/06/07 14:09:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
(ORD,108984)
(ATL,106898)
(DFW,70657)
(DEN,63003)
(LAX,59969)
Top 5 most visited destinations18/06/07 14:09:38 INFO Executor: Finished task 0.0 in stage 5.0
18/06/07 14:09:38 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 25) in 47 ms on loca
18/06/07 14:09:38 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed
```

Problem Statement 2

Which month has seen the most number of cancellations due to bad weather?

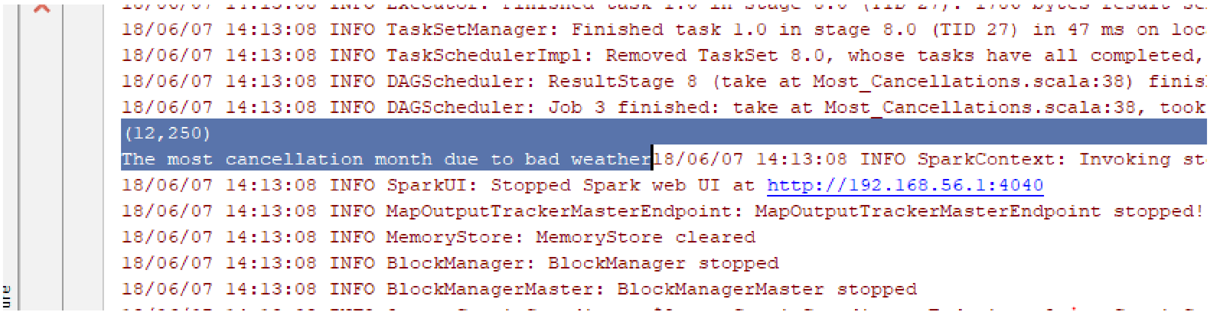


```scala
package Session28_Assignment1

import org.apache.spark.sql.SparkSession

object Most_Cancellations {

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

    println("Session 28 assignment problem 2 !!!")

    // Use new SparkSession interface in Spark
    val spark = SparkSession
      .builder()
      .master( master = "local[*]")
      .appName( name = "Assignment")
      .config("spark.some.config.option","some-value")
      .getOrCreate()

    // load the dataset using the textFile method.
    val delayed_Flights_data_with_header = spark.sparkContext.textFile( path = "C:\\Users\\Bhaskar\\Desktop\\AcadGild\\AcadgildSessions\\Sessio

    //creating a variable header, which holds the first line of the dataset, in our data set Sports_data.txt the first line is a header line.
    val header = delayed_Flights_data_with_header.first()

    //filter the header line from the dataset using the filter RDD
    val delayed_Flights_data = delayed_Flights_data_with_header.filter(row => row != header)
```



```scala
    //creating a variable header, which holds the first line of the dataset, in our data set Sports_data.txt the first line is a header line.
    val header = delayed_Flights_data_with_header.first()

    //filter the header line from the dataset using the filter RDD
    val delayed_Flights_data = delayed_Flights_data_with_header.filter(row => row != header)

    // filter the null records from delayed flight data
    val filter_null_values = delayed_Flights_data.map(x => x.split( regex = ",")).filter(x => x!= null)

    //filter column cancelled with value 1 or yes and cancellation code for weather "B" and map month column as key
    val a = filter_null_values.filter(x => ((x(22).equals("1"))&&(x(23).equals("B")))).map(x => (x(2),1))

    //use reduce by key to calculate total no. of cancellation for each month and sorted in descending order
    val b = a.reduceByKey(_+_).map(x => (x._2,x._1)).sortByKey( ascending = false)

    // print the top cancellation month
    val c = b.map(x => (x._2,x._1)).take( num = 1).foreach(println)

    print("The most cancellation month due to bad weather")
  }
}
```

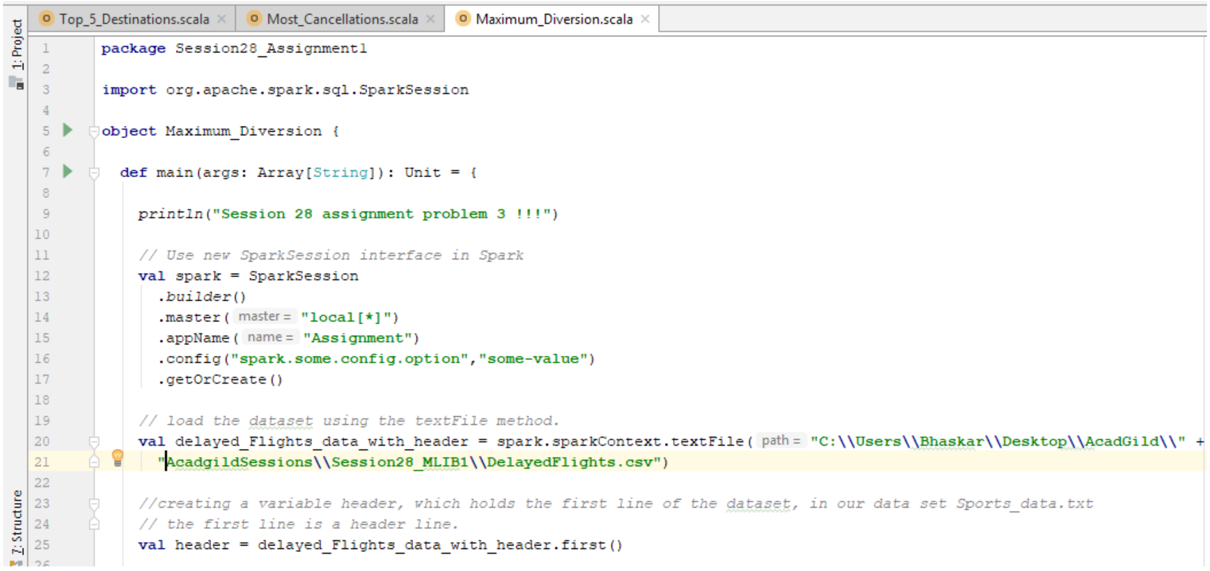
Output:



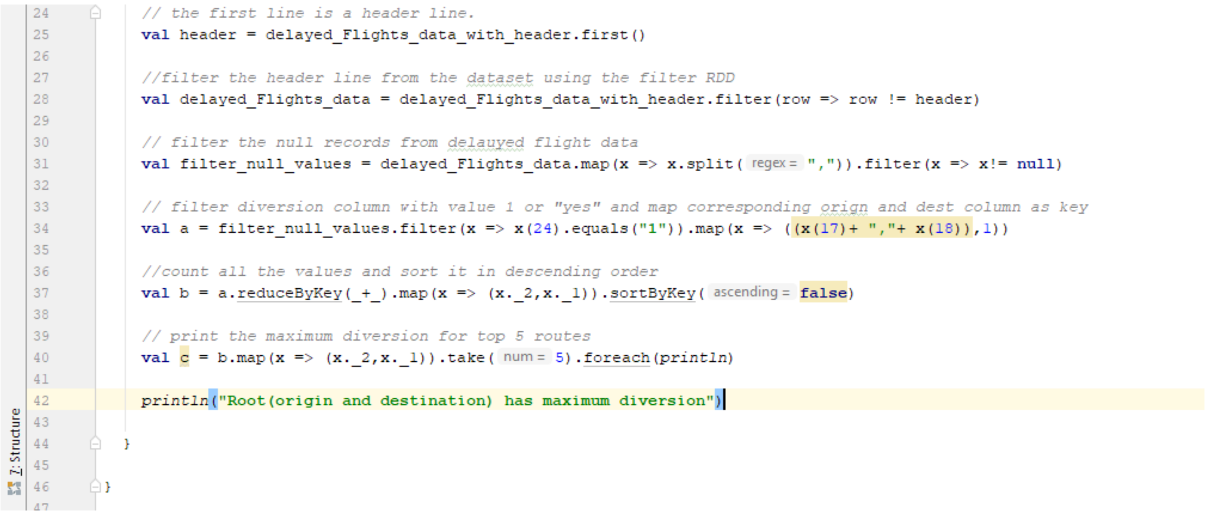
```
18/06/07 14:13:08 INFO TaskSetManager: Finished task 1.0 in stage 8.0 (TID 27) in 47 ms on loc
18/06/07 14:13:08 INFO TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have all completed,
18/06/07 14:13:08 INFO DAGScheduler: ResultStage 8 (take at Most_Cancellations.scala:38) finis
18/06/07 14:13:08 INFO DAGScheduler: Job 3 finished: take at Most_Cancellations.scala:38, took
(12,250)
The most cancellation month due to bad weather18/06/07 14:13:08 INFO SparkContext: Invoking st
18/06/07 14:13:08 INFO SparkUI: Stopped Spark web UI at http://192.168.56.1:4040
18/06/07 14:13:08 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/06/07 14:13:08 INFO MemoryStore: MemoryStore cleared
18/06/07 14:13:08 INFO BlockManager: BlockManager stopped
18/06/07 14:13:08 INFO BlockManagerMaster: BlockManagerMaster stopped
```

Problem Statement 3

Which route (origin & destination) has seen the maximum diversion?

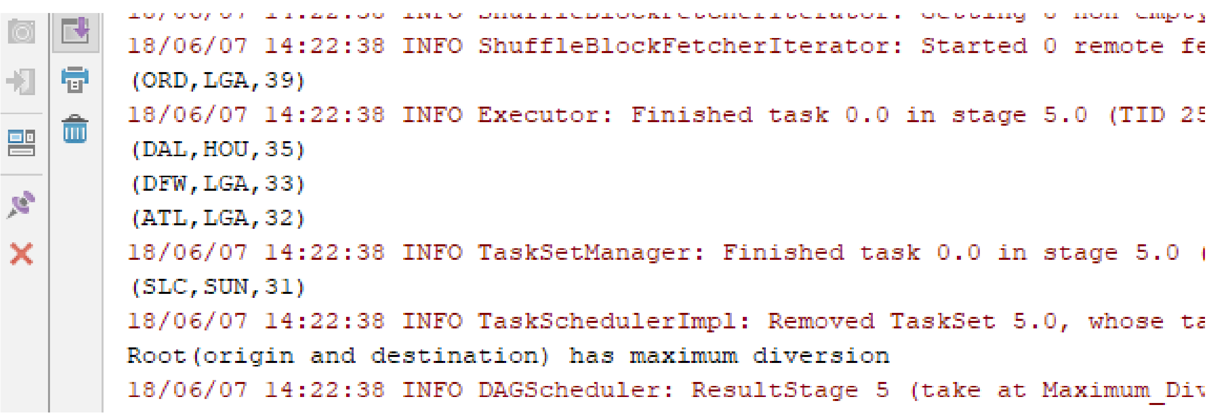


```scala
package Session28_Assignment1

import org.apache.spark.sql.SparkSession

object Maximum_Diversion {

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

    println("Session 28 assignment problem 3 !!!")

    // Use new SparkSession interface in Spark
    val spark = SparkSession
      .builder()
      .master( master = "local[*]")
      .appName( name = "Assignment")
      .config("spark.some.config.option","some-value")
      .getOrCreate()

    // load the dataset using the textFile method.
    val delayed_Flights_data_with_header = spark.sparkContext.textFile( path = "C:\\Users\\Bhaskar\\Desktop\\AcadGild\\" +
      "AcadgildSessions\\Session28_MLIB1\\DelayedFlights.csv")

    //creating a variable header, which holds the first line of the dataset, in our data set Sports_data.txt
    // the first line is a header line.
    val header = delayed_Flights_data_with_header.first()
```



```scala
    // the first line is a header line.
    val header = delayed_Flights_data_with_header.first()

    //filter the header line from the dataset using the filter RDD
    val delayed_Flights_data = delayed_Flights_data_with_header.filter(row => row != header)

    // filter the null records from delauyed flight data
    val filter_null_values = delayed_Flights_data.map(x => x.split( regex = ",")).filter(x => x!= null)

    // filter diversion column with value 1 or "yes" and map corresponding orign and dest column as key
    val a = filter_null_values.filter(x => x(24).equals("1")).map(x => ((x(17)+ ","+ x(18)),1))

    //count all the values and sort it in descending order
    val b = a.reduceByKey(_+_).map(x => (x._2,x._1)).sortByKey( ascending = false)

    // print the maximum diversion for top 5 routes
    val c = b.map(x => (x._2,x._1)).take( num = 5).foreach(println)

    println("Root(origin and destination) has maximum diversion")

  }

}
```

Output:



```
18/06/07 14:22:38 INFO ShuffleBlockFetcherIterator: Getting 0 non-empty
18/06/07 14:22:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fe
(ORD,LGA,39)
18/06/07 14:22:38 INFO Executor: Finished task 0.0 in stage 5.0 (TID 25
(DAL,HOU,35)
(DFW,LGA,33)
(ATL,LGA,32)
18/06/07 14:22:38 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (
(SLC,SUN,31)
18/06/07 14:22:38 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose ta
Root(origin and destination) has maximum diversion
18/06/07 14:22:38 INFO DAGScheduler: ResultStage 5 (take at Maximum_Div
```