```
// Assignment 4 - Domas Budrys

//Question 1
//Added headers manually to each file

//Alabama
val alDF = spark.read.option("header", "true").option("inferSchema",
"true").csv("/Users/domo/Desktop/Domas-A4/AL.TXT")

//Georgia
val gaDF = spark.read.option("header", "true").option("inferSchema",
"true").csv("/Users/domo/Desktop/Domas-A4/GA.TXT")

//Kentucky
val kyDF = spark.read.option("header", "true").option("inferSchema",
"true").csv("/Users/domo/Desktop/Domas-A4/KY.TXT")

//Tennessee
val tnDF = spark.read.option("header", "true").option("inferSchema",
"true").csv("/Users/domo/Desktop/Domas-A4/TN.TXT")




//Question 2
val namesDF = alDF.union(gaDF).union(kyDF).union(tnDF)



//Question 3
val pivotQ3 = namesDF.groupBy("Sex").pivot("State").sum()
pivotQ3.select("Sex", "AL_sum(Count)", "GA_sum(Count)",
"KY_sum(Count)", "TN_sum(Count)").show()



//Question 4
val pivotQ4 = namesDF.groupBy("State").pivot("State").sum()
pivotQ4.select("State", "AL_sum(Count)", "GA_sum(Count)",
"KY_sum(Count)", "TN_sum(Count)").show()



//Question 5

import org.apache.spark.sql.functions._
val pivotQ5 = namesDF.groupBy("Name").pivot("State").sum()
val pivotQ5Count = pivotQ5.select("Name", "AL_sum(Count)",
"GA_sum(Count)", "KY_sum(Count)",
"TN_sum(Count)").withColumn("TotalCount", pivotQ5("AL_sum(Count)") +
pivotQ5("GA_sum(Count)")+ pivotQ5("KY_sum(Count)") +
pivotQ5("TN_sum(Count)"))
pivotQ5Count.select("Name",
"TotalCount").orderBy(desc("TotalCount")).limit(5).show()



//Question 6

val pivotQ6 = namesDF.groupBy("Name", "Sex").pivot("State").sum()
val pivotQ6Count = pivotQ6.select("Name", "Sex", "AL_sum(Count)",
"GA_sum(Count)", "KY_sum(Count)",
"TN_sum(Count)").withColumn("TotalCount", pivotQ6("AL_sum(Count)") +
pivotQ6("GA_sum(Count)")+ pivotQ6("KY_sum(Count)") +
pivotQ6("TN_sum(Count)"))
pivotQ6Count.where("Sex = 'F'").select("Name", "Sex",
"TotalCount").orderBy(desc("TotalCount")).limit(5).show()




//Question 7

val pivotQ7 = namesDF.groupBy("Name", "Sex").pivot("State").sum()
val pivotQ7Count = pivotQ7.select("Name", "Sex", "AL_sum(Count)",
"GA_sum(Count)", "KY_sum(Count)",
"TN_sum(Count)").withColumn("TotalCount", pivotQ7("AL_sum(Count)") +
pivotQ7("GA_sum(Count)")+ pivotQ7("KY_sum(Count)") +
pivotQ7("TN_sum(Count)"))
pivotQ7Count.where("Sex = 'M'").select("Name", "Sex",
"TotalCount").orderBy(desc("TotalCount")).limit(5).show()
```