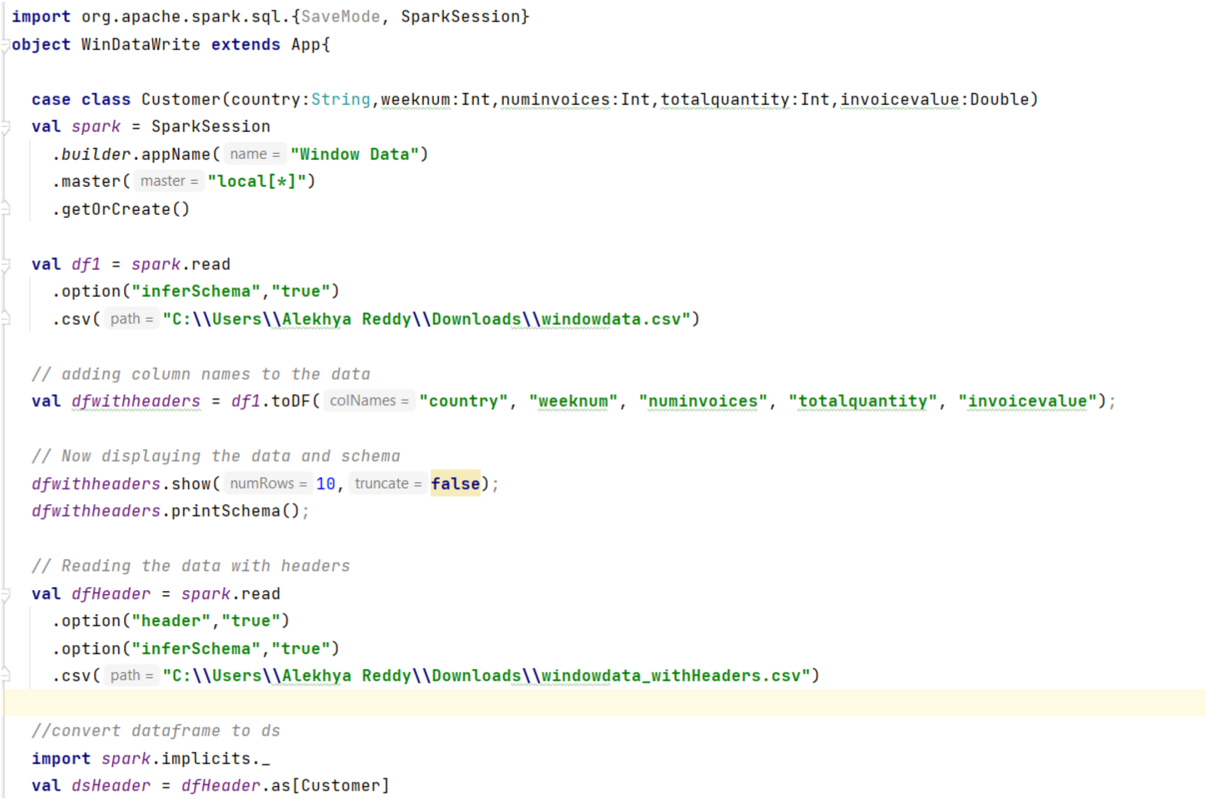
Question:

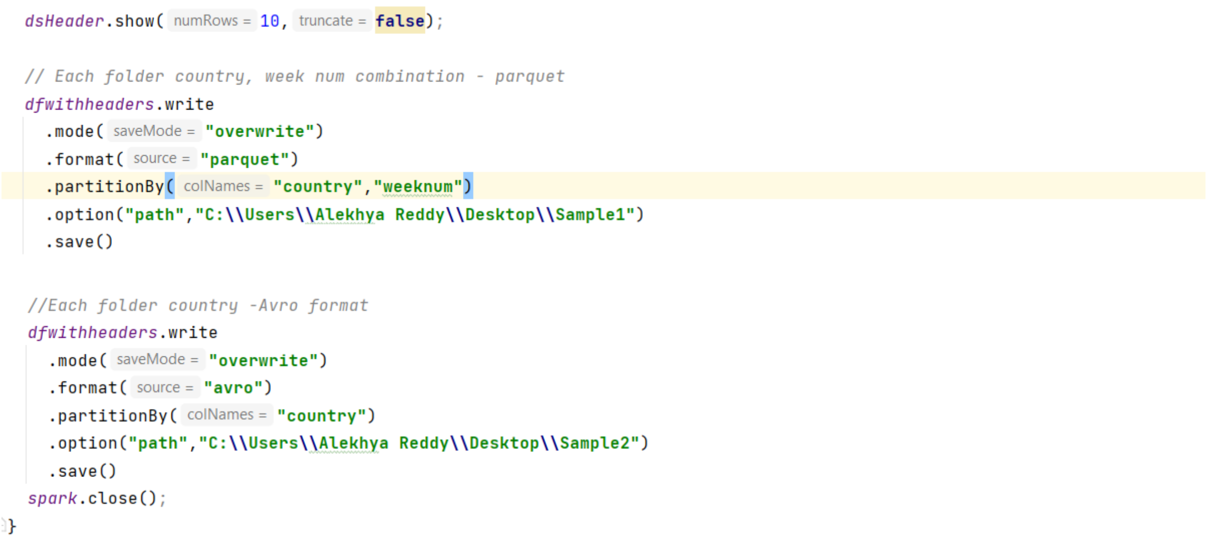
We have a file windowdata.csv and the field names are country, weeknum, numinvoices, totalquantity, invoicevalue

Step 1: create spark session
Step 2: set the logging level to error
Step 3: Using the standard dataframe reader API load the file and create a dataframe.
Step 4: Use the standard dataframe writer api to save it in parquet format. While saving make sure data is stored where we should have a folder for each country, weeknum (combination)
Step 5: Also use the dataframe write api to save the data in Avro format. While saving make sure data is stored where we should have a folder for each country.
Step 6: Apply header
Step 7: Convert dataframe to dataset(Specific type)

Program:

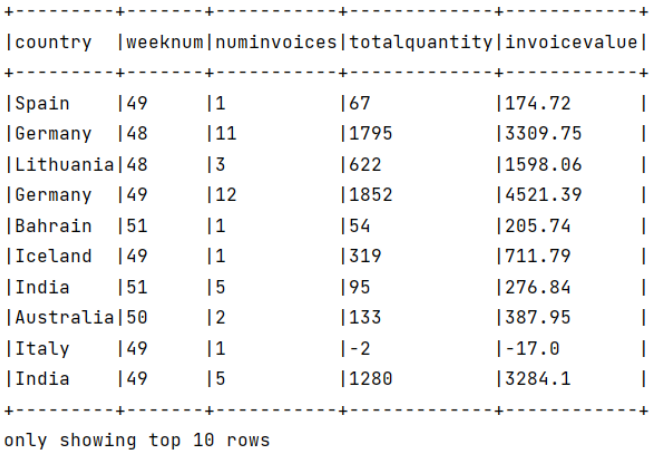


```
import org.apache.spark.sql.{SaveMode, SparkSession}
object WinDataWrite extends App{

  case class Customer(country:String,weeknum:Int,numinvoices:Int,totalquantity:Int,invoicevalue:Double)
  val spark = SparkSession
    .builder.appName( name = "Window Data")
    .master( master = "local[*]")
    .getOrCreate()

  val df1 = spark.read
    .option("inferSchema","true")
    .csv( path = "C:\\Users\\Alekhya Reddy\\Downloads\\windowdata.csv")

  // adding column names to the data
  val dfwithheaders = df1.toDF( colNames = "country", "weeknum", "numinvoices", "totalquantity", "invoicevalue");

  // Now displaying the data and schema
  dfwithheaders.show( numRows = 10, truncate = false);
  dfwithheaders.printSchema();

  // Reading the data with headers
  val dfHeader = spark.read
    .option("header","true")
    .option("inferSchema","true")
    .csv( path = "C:\\Users\\Alekhya Reddy\\Downloads\\windowdata_withHeaders.csv")

  //convert dataframe to ds
  import spark.implicits._
  val dsHeader = dfHeader.as[Customer]
```

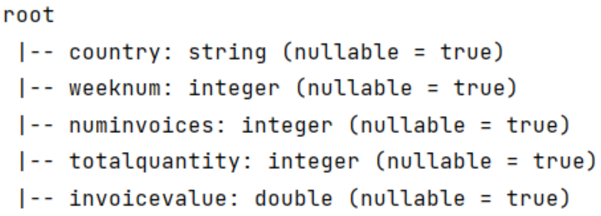


```
dsHeader.show( numRows = 10, truncate = false);

// Each folder country, week num combination - parquet
dfwithheaders.write
  .mode( saveMode = "overwrite")
  .format( source = "parquet")
  .partitionBy( colNames = "country","weeknum")
  .option("path","C:\\Users\\Alekhya Reddy\\Desktop\\Sample1")
  .save()

//Each folder country -Avro format
dfwithheaders.write
  .mode( saveMode = "overwrite")
  .format( source = "avro")
  .partitionBy( colNames = "country")
  .option("path","C:\\Users\\Alekhya Reddy\\Desktop\\Sample2")
  .save()
spark.close();
}
```

O/p:

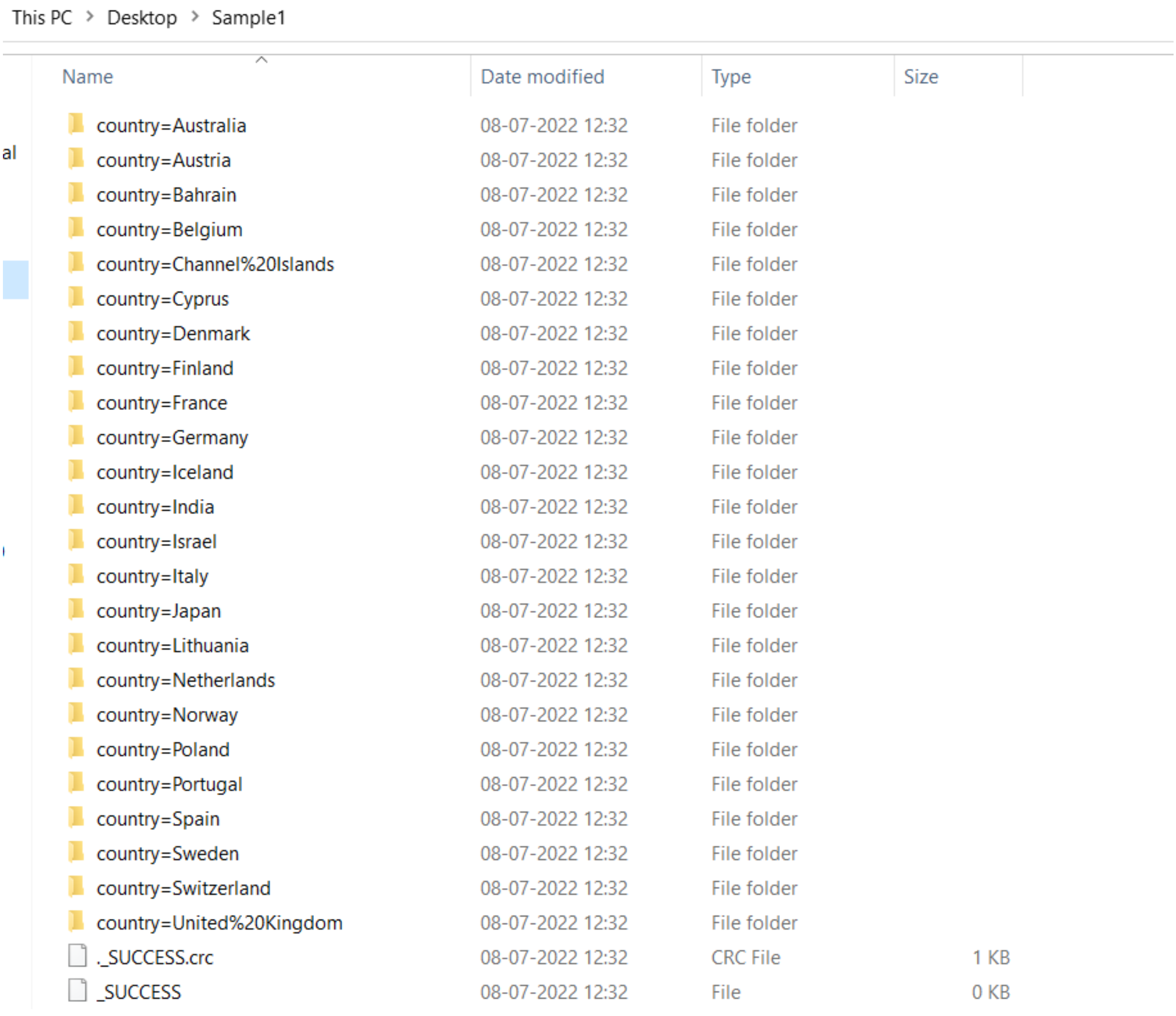


```
+---------+-------+-----------+-------------+------------+
|country  |weeknum|numinvoices|totalquantity|invoicevalue|
+---------+-------+-----------+-------------+------------+
|Spain    |49     |1          |67           |174.72      |
|Germany  |48     |11         |1795         |3309.75     |
|Lithuania|48     |3          |622          |1598.06     |
|Germany  |49     |12         |1852         |4521.39     |
|Bahrain  |51     |1          |54           |205.74      |
|Iceland  |49     |1          |319          |711.79      |
|India    |51     |5          |95           |276.84      |
|Australia|50     |2          |133          |387.95      |
|Italy    |49     |1          |-2           |-17.0       |
|India    |49     |5          |1280         |3284.1      |
+---------+-------+-----------+-------------+------------+
only showing top 10 rows
```



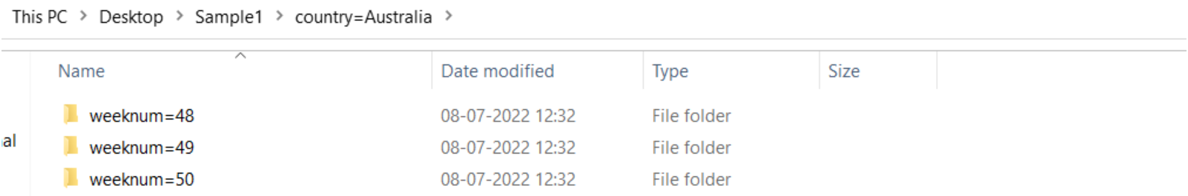
```
root
 |-- country: string (nullable = true)
 |-- weeknum: integer (nullable = true)
 |-- numinvoices: integer (nullable = true)
 |-- totalquantity: integer (nullable = true)
 |-- invoicevalue: double (nullable = true)
```

Sample1 -PARQUET Format



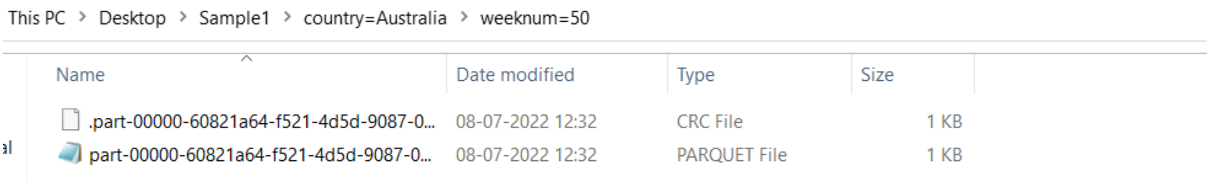
This PC › Desktop › Sample1

Name	Date modified	Type	Size
country=Australia	08-07-2022 12:32	File folder	
country=Austria	08-07-2022 12:32	File folder	
country=Bahrain	08-07-2022 12:32	File folder	
country=Belgium	08-07-2022 12:32	File folder	
country=Channel%20Islands	08-07-2022 12:32	File folder	
country=Cyprus	08-07-2022 12:32	File folder	
country=Denmark	08-07-2022 12:32	File folder	
country=Finland	08-07-2022 12:32	File folder	
country=France	08-07-2022 12:32	File folder	
country=Germany	08-07-2022 12:32	File folder	
country=Iceland	08-07-2022 12:32	File folder	
country=India	08-07-2022 12:32	File folder	
country=Israel	08-07-2022 12:32	File folder	
country=Italy	08-07-2022 12:32	File folder	
country=Japan	08-07-2022 12:32	File folder	
country=Lithuania	08-07-2022 12:32	File folder	
country=Netherlands	08-07-2022 12:32	File folder	
country=Norway	08-07-2022 12:32	File folder	
country=Poland	08-07-2022 12:32	File folder	
country=Portugal	08-07-2022 12:32	File folder	
country=Spain	08-07-2022 12:32	File folder	
country=Sweden	08-07-2022 12:32	File folder	
country=Switzerland	08-07-2022 12:32	File folder	
country=United%20Kingdom	08-07-2022 12:32	File folder	
._SUCCESS.crc	08-07-2022 12:32	CRC File	1 KB
_SUCCESS	08-07-2022 12:32	File	0 KB



This PC › Desktop › Sample1 › country=Australia ›

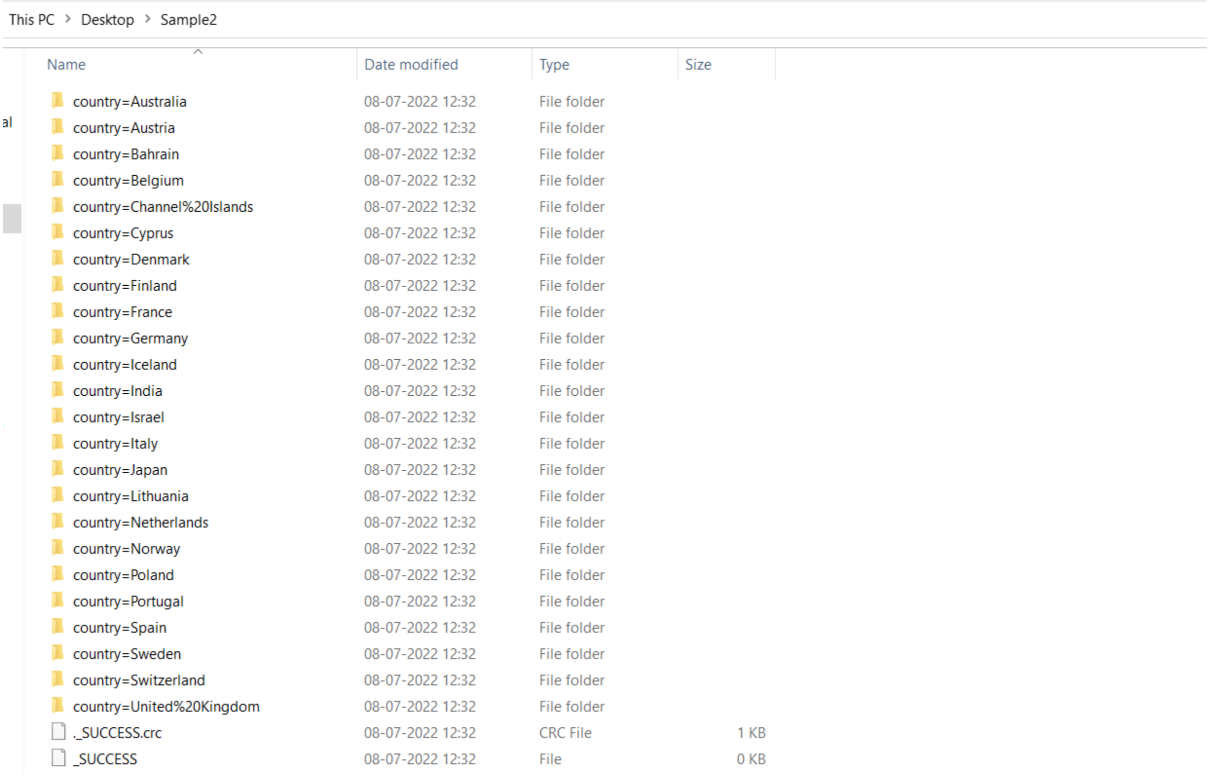
Name	Date modified	Type	Size
weeknum=48	08-07-2022 12:32	File folder	
weeknum=49	08-07-2022 12:32	File folder	
weeknum=50	08-07-2022 12:32	File folder	



This PC › Desktop › Sample1 › country=Australia › weeknum=50

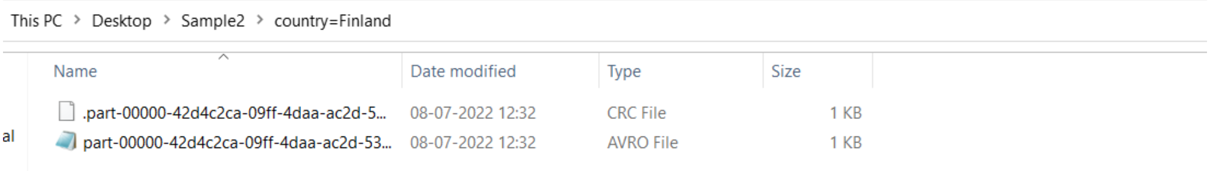
Name	Date modified	Type	Size
.part-00000-60821a64-f521-4d5d-9087-0...	08-07-2022 12:32	CRC File	1 KB
part-00000-60821a64-f521-4d5d-9087-0...	08-07-2022 12:32	PARQUET File	1 KB

Sample2 – AVRO Format



This PC › Desktop › Sample2

Name	Date modified	Type	Size
country=Australia	08-07-2022 12:32	File folder	
country=Austria	08-07-2022 12:32	File folder	
country=Bahrain	08-07-2022 12:32	File folder	
country=Belgium	08-07-2022 12:32	File folder	
country=Channel%20Islands	08-07-2022 12:32	File folder	
country=Cyprus	08-07-2022 12:32	File folder	
country=Denmark	08-07-2022 12:32	File folder	
country=Finland	08-07-2022 12:32	File folder	
country=France	08-07-2022 12:32	File folder	
country=Germany	08-07-2022 12:32	File folder	
country=Iceland	08-07-2022 12:32	File folder	
country=India	08-07-2022 12:32	File folder	
country=Israel	08-07-2022 12:32	File folder	
country=Italy	08-07-2022 12:32	File folder	
country=Japan	08-07-2022 12:32	File folder	
country=Lithuania	08-07-2022 12:32	File folder	
country=Netherlands	08-07-2022 12:32	File folder	
country=Norway	08-07-2022 12:32	File folder	
country=Poland	08-07-2022 12:32	File folder	
country=Portugal	08-07-2022 12:32	File folder	
country=Spain	08-07-2022 12:32	File folder	
country=Sweden	08-07-2022 12:32	File folder	
country=Switzerland	08-07-2022 12:32	File folder	
country=United%20Kingdom	08-07-2022 12:32	File folder	
._SUCCESS.crc	08-07-2022 12:32	CRC File	1 KB
_SUCCESS	08-07-2022 12:32	File	0 KB



This PC › Desktop › Sample2 › country=Finland

Name	Date modified	Type	Size
.part-00000-42d4c2ca-09ff-4daa-ac2d-5...	08-07-2022 12:32	CRC File	1 KB
part-00000-42d4c2ca-09ff-4daa-ac2d-53...	08-07-2022 12:32	AVRO File	1 KB