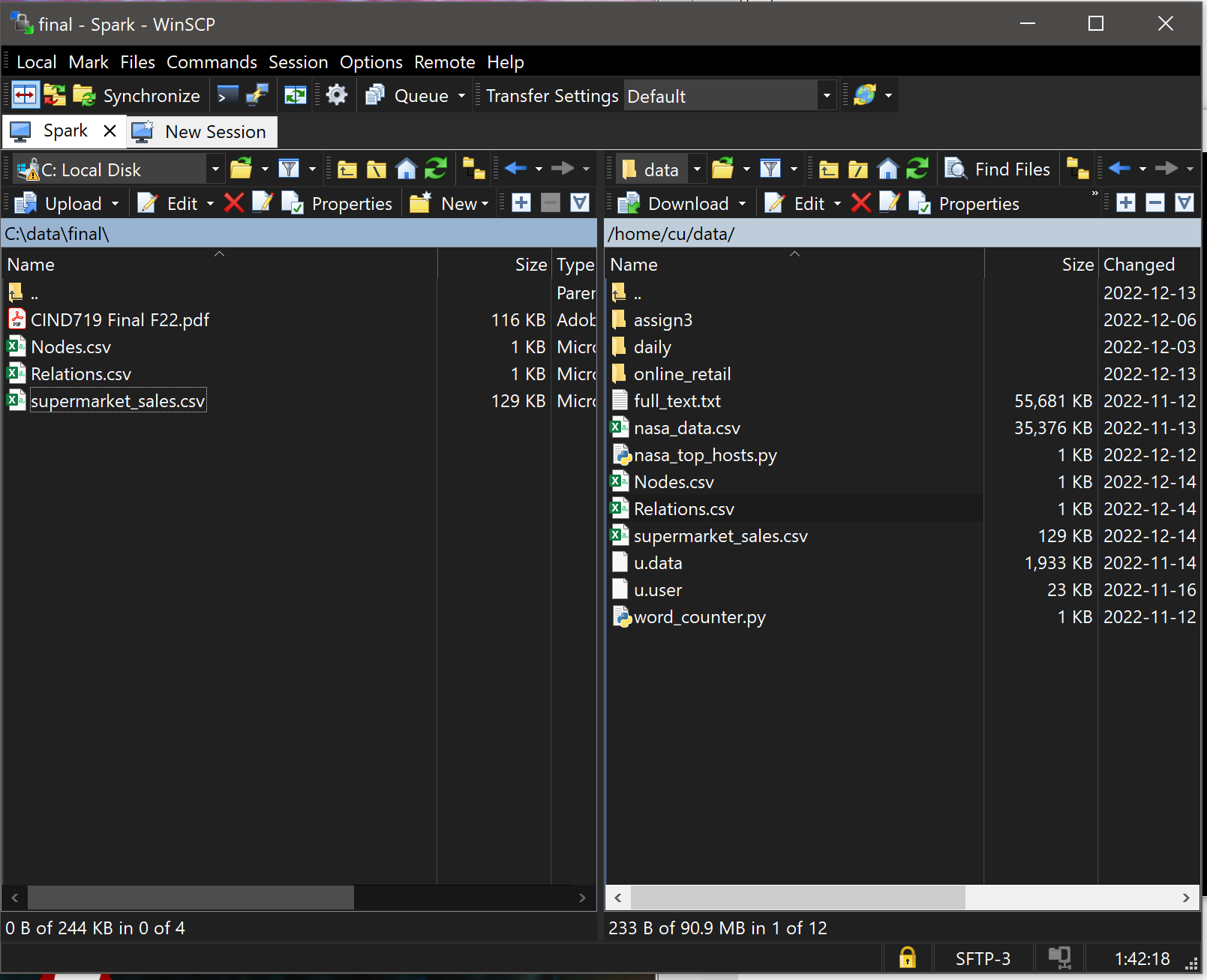
**CIND 719: Big Data Analytics Tools**

**Final Exam**

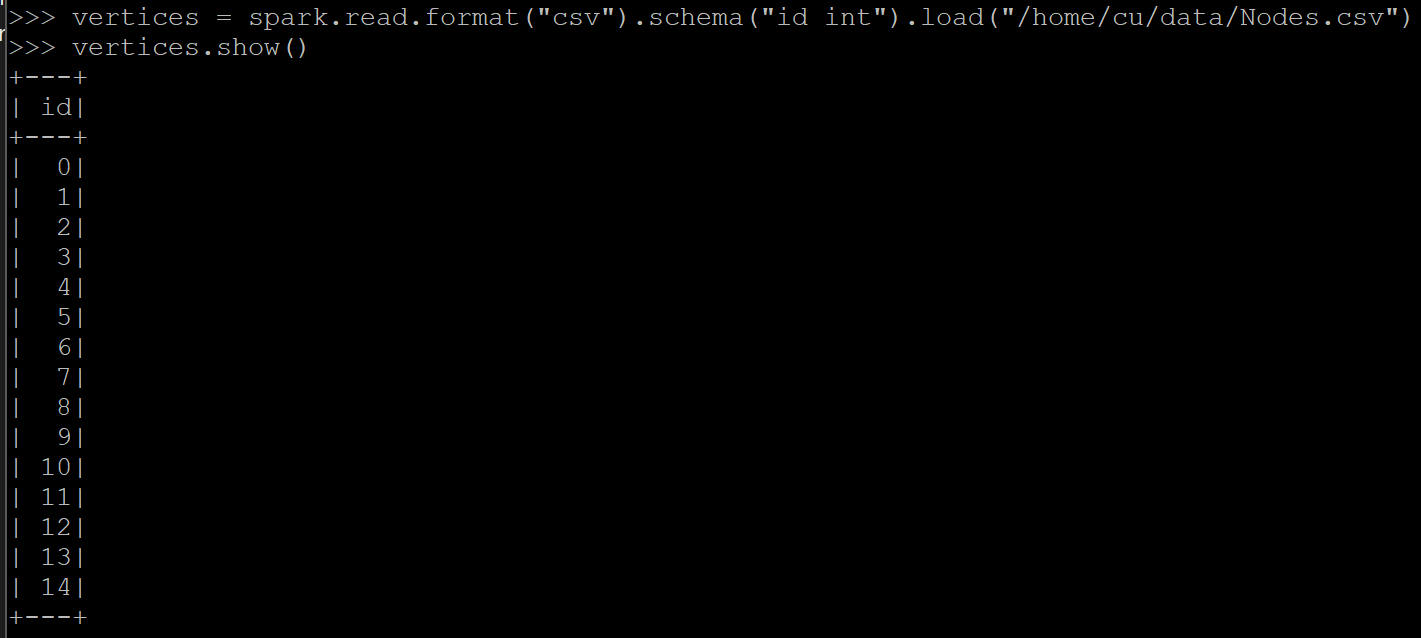
Importing data files: Using WINSCP to transfer the files into the /home/cu/data folder4



**Question 1**

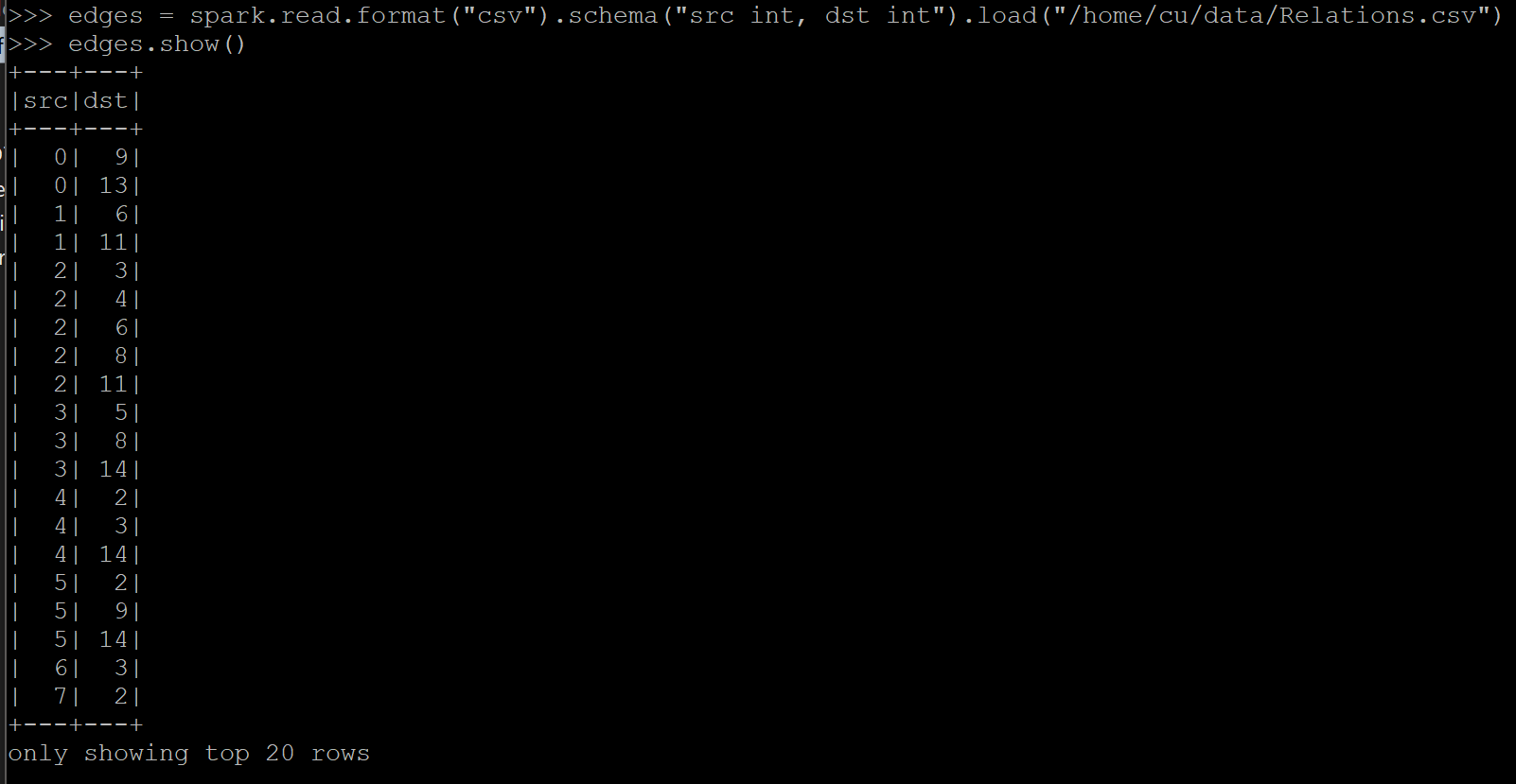
1. >>> vertices = spark.read.format("csv").schema("id int").load("/home/cu/data/Nodes.csv")

>>> vertices.show()



>>> edges = spark.read.format("csv").schema("src int, dst int").load("/home/cu/data/Relations.csv")

>>> edges.show()

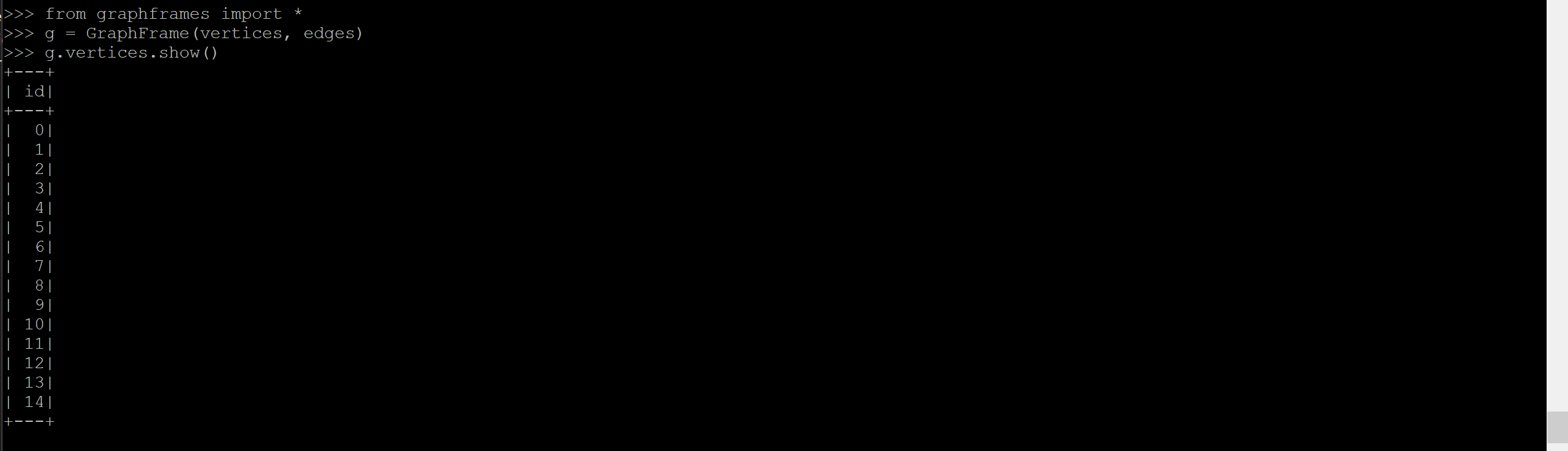


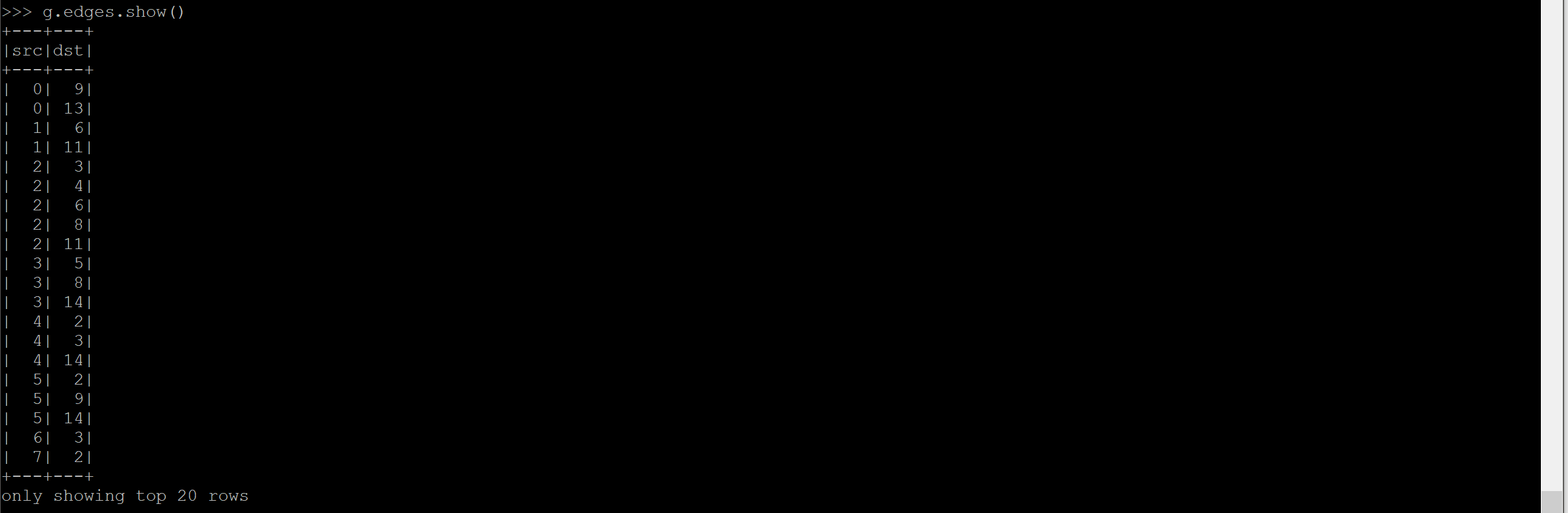
1. >>> from graphframes import \*

>>> g = GraphFrame(vertices, edges)

>>> g.vertices.show()

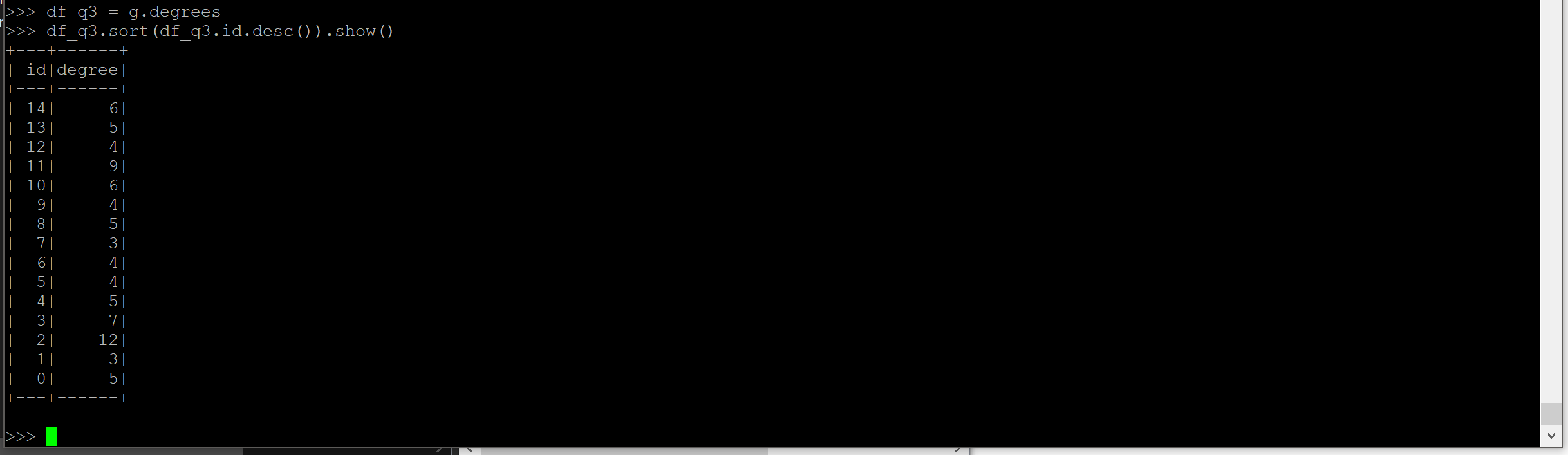
>>> g.edges.show()



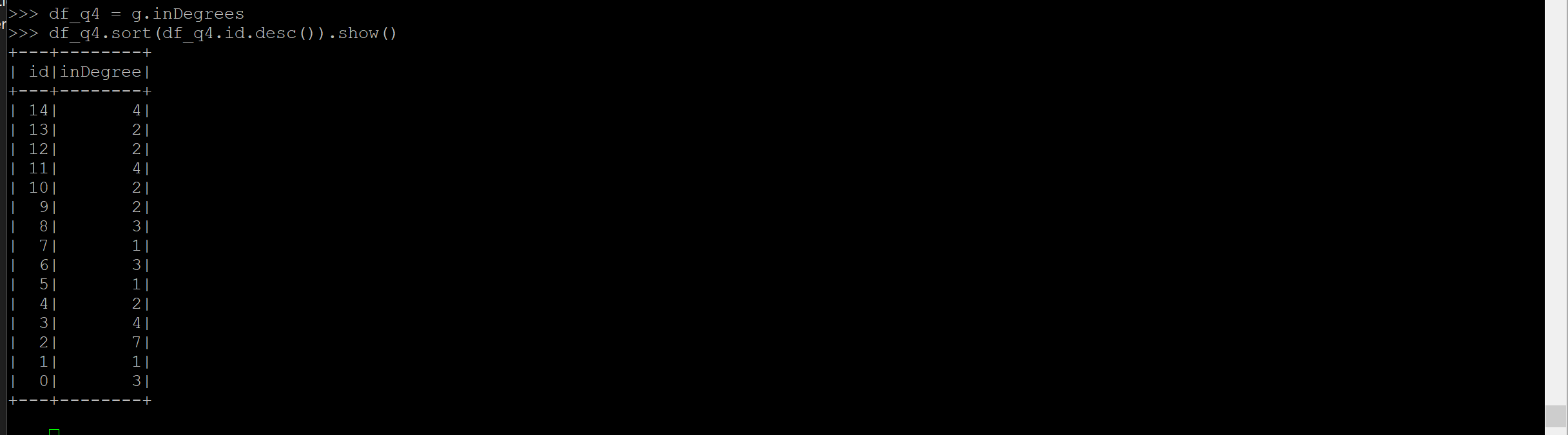


1. >>> df\_q3 = g.degrees

>>> df\_q3.sort(df\_q3.id.desc()).show()



1. >>> df\_q4 = g.inDegrees

>>> df\_q4.sort(df\_q4.id.desc()).show()  


1. >>> motifs = g.find("(a)-[]->(b);(b)-[]->(c);(c)-[]->(a)")

>>> motifs.show()

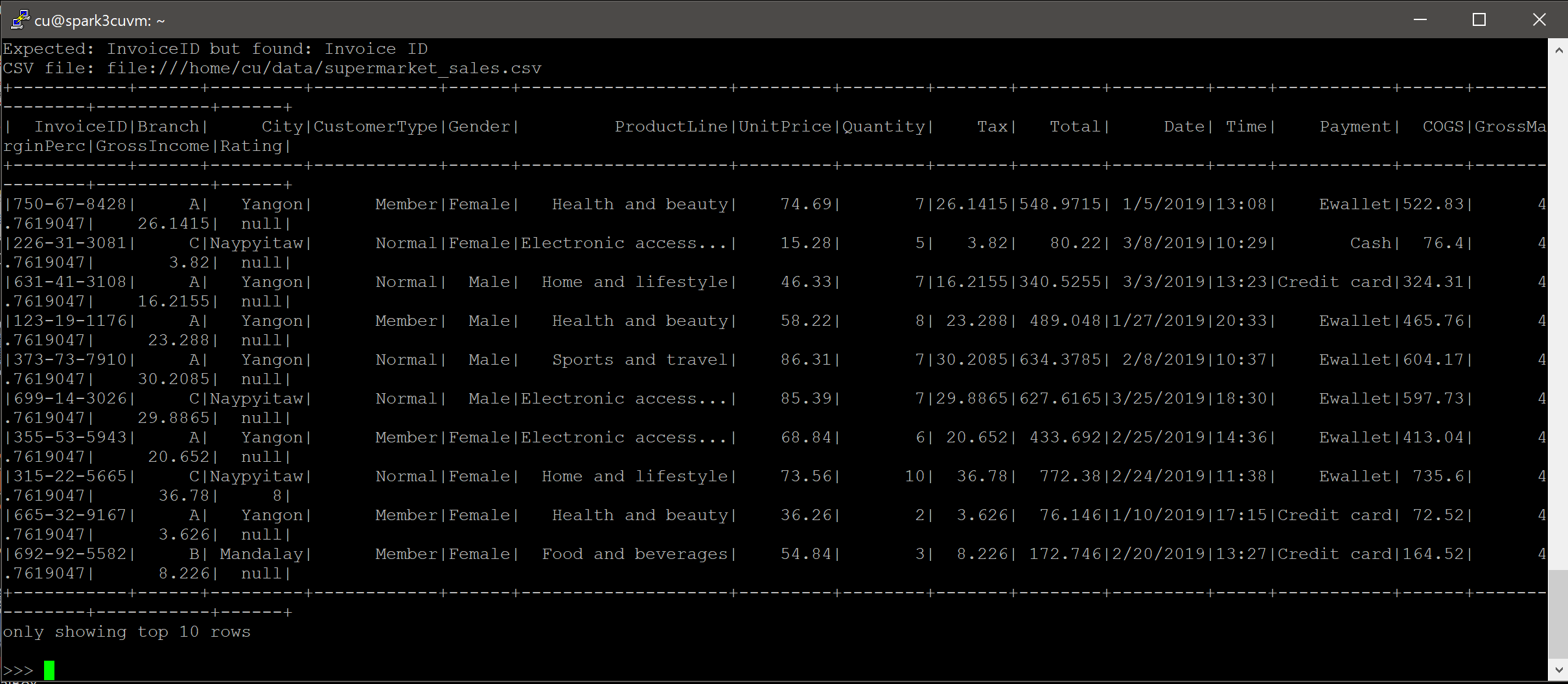


**Question 2**

1. >>> schema = "InvoiceID string," " Branch string," " City string," "CustomerType string," "Gender string," "ProductLine string," "UnitPrice float," "Quantity int," "Tax float," "Total float," "Date string," "Time string," "Payment string," "COGS float," "GrossMarginPerc float," "GrossIncome float," "Rating int"

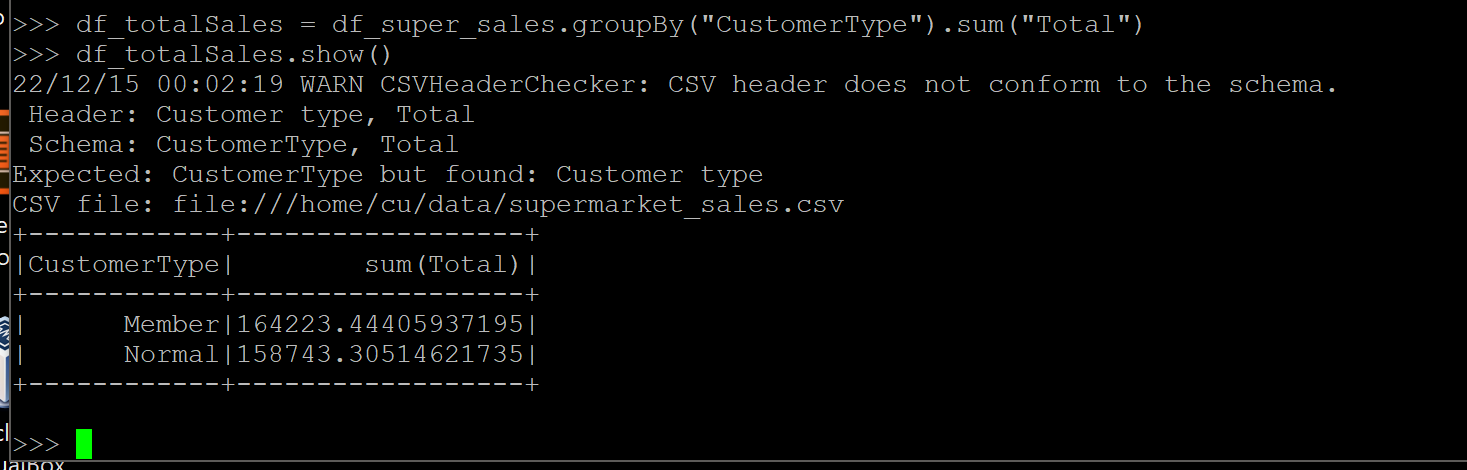
>>> df\_super\_sales = spark.read.format("csv").schema(schema).option("delimiter", ",").option("header", "true").load("/home/cu/data/supermarket\_sales.csv")

>>> df\_super\_sales.show(10)



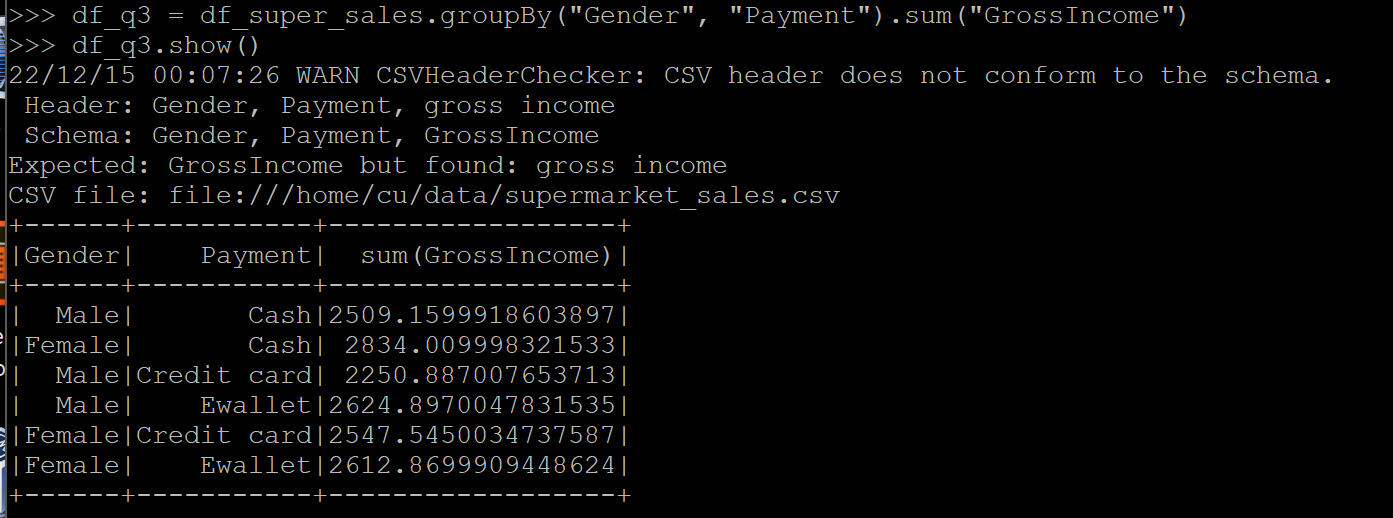
1. >>> df\_totalSales = df\_super\_sales.groupBy("CustomerType").sum("Total")

>>> df\_totalSales.show()



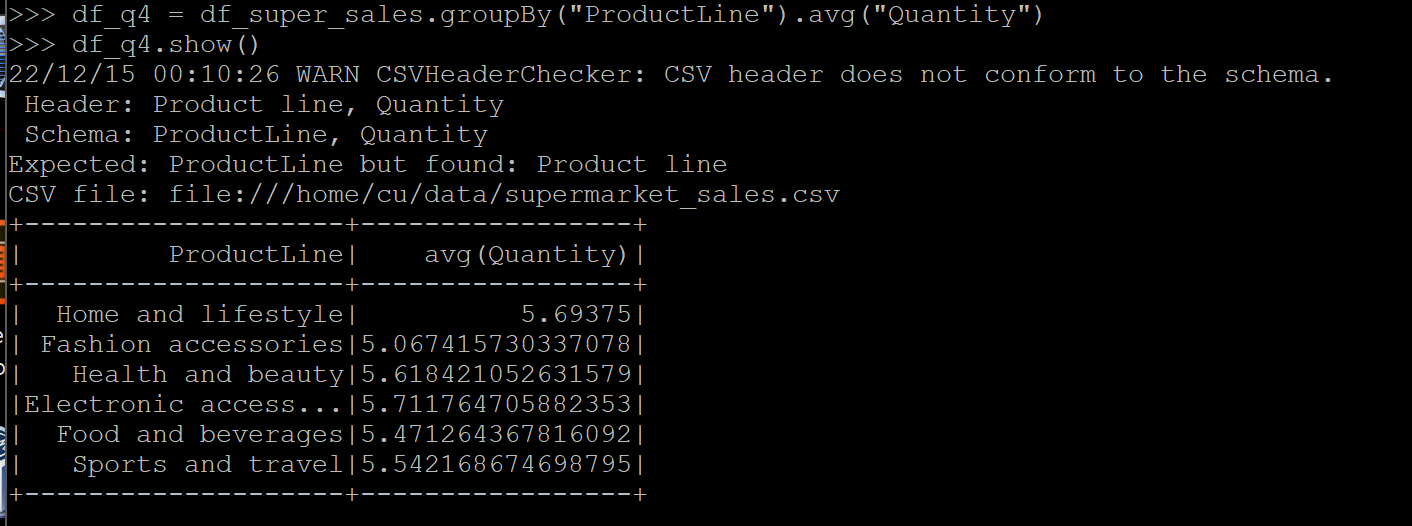
1. >>> df\_q3 = df\_super\_sales.groupBy("Gender", "Payment").sum("GrossIncome")

>>> df\_q3.show()



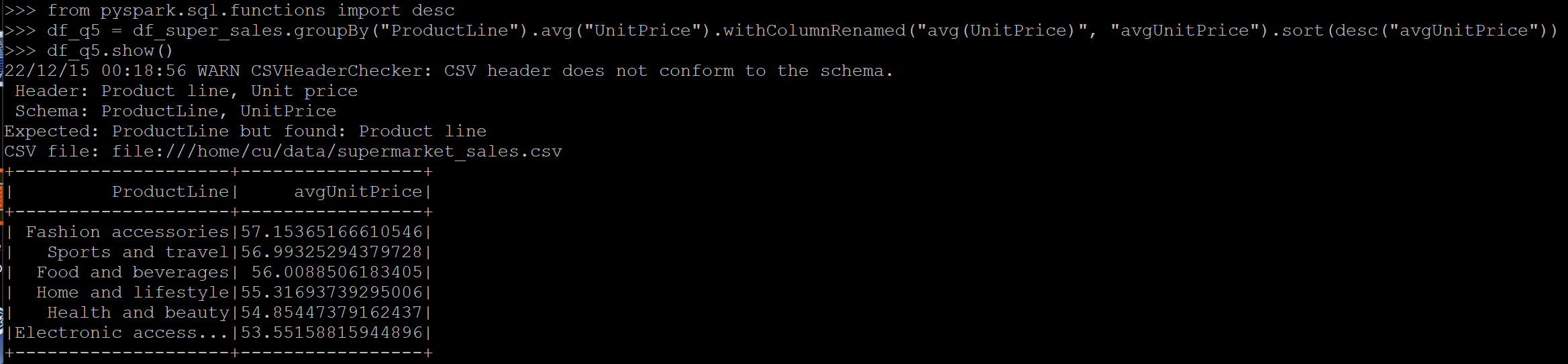
1. >>> df\_q4 = df\_super\_sales.groupBy("ProductLine").avg("Quantity")

>>> df\_q4.show()



1. >>> df\_q5 = df\_super\_sales.groupBy("ProductLine").avg("UnitPrice").withColumnRenamed("avg(UnitPrice)", "avgUnitPrice").sort(desc("avgUnitPrice"))

>>> df\_q5.show()



1. >>> df\_q6 = df\_super\_sales.groupBy("Gender", "ProductLine").sum("Quantity").withColumnRenamed("sum(Quantity)", "sumQuantity").sort(desc("sumQuantity"))

>>> df\_q6.show()