Exercise 2: CSV File Import in Databricks

path = "/databricks-datasets/online\_retail/data-001/data.csv"

Question 1: How to load a csv file?

path = "/databricks-datasets/online\_retail/data-001/data.csv"

df = spark.read.load(path, format='com.databricks.spark.csv', header='true', inferSchema='true')

Question 2: Show 20 rows

display(df)

Question 3: Show me number of rows

Question 4: Show me first rows

Question 5: Show me first 10 rows

Question 6: Show DataFrame Schema

df.printSchema()

Question 7: Show “Country” column value

df.select("Country").show()

Question 8: **Remove Duplicates from Column and Sort**

display(

df .select("Country")

.distinct()

.orderBy("Country"))

Question 9: **Calculate Order Totals**

display(

df

.select(df["InvoiceNo"],df["UnitPrice"]\*df["Quantity"])

.groupBy("InvoiceNo")

.sum() )

Question 10: **Inspect Results with Filter of** InvoiceNo=536596

df.filter(df["InvoiceNo"]==536596).show()

Question 11: **Show Top 10 Products in the UK with sum of total sells**

display(

df .select(df["Country"], df["Description"],(df["UnitPrice"]\*df["Quantity"]).alias("Total")) .groupBy("Country", "Description")

.sum()

.filter(df["Country"]=="United Kingdom")

.sort("sum(Total)", ascending=**False**)

.limit(10) )

Question 12: **calculate product sales by country**

r1 = df.select(df["Country"], df["Description"],(df["UnitPrice"]\*df["Quantity"]).alias("Total")) display(r1)

Question 13: **save result as table**

r1.write.saveAsTable("product\_sales\_by\_country")