Group Members Usman Zaman 2211-032-KHI-DEG & Syed Wasay Waseem 2211-030-KHI-DEG

In [1]:
```
from pyspark.sql import SparkSession
from pyspark.sql.types import IntegerType
```

In [2]:
```
scSpark = SparkSession.builder.appName("Assignment_5.1").getOrCreate()
```

Reading datasets and creating a joined dataframe

In [3]:
```
df_1 = scSpark.read.csv("data_2/transactions*.csv", header=True)
df_2 = scSpark.read.csv("data_2/products.csv", header=True)
df_3 = scSpark.read.csv("data_2/customers.csv", header=True)
```

In [4]:
```
merged_df=df_1.join(df_2,"ProductId","inner")
```

In [5]:
```
merged_df.show()
```

```
+---------+-------+-------------+----------+--------+-------------------+-------------+--------+---------+
|ProductId|StoreId|TransactionId|CustomerId|Quantity|    TransactionTime|         Name|Category|UnitPrice|
+---------+-------+-------------+----------+--------+-------------------+-------------+--------+---------+
|        3|      3|          454|        35|       3|2022-12-23 17:36:11|  Blue Shorts|  Shorts|   118.88|
|        9|      3|          524|        37|      11|2022-12-23 22:02:51|Green Sandals|   Shoes|   137.53|
|        3|      3|          562|         4|       4|2022-12-23 02:51:50|  Blue Shorts|  Shorts|   118.88|
|       14|      3|          581|        35|      56|2022-12-23 17:05:54|  Red t-shirt|T-Shirts|   121.58|
|       15|      3|          200|        34|      24|2022-12-23 07:15:01|White t-shirt|T-Shirts|   131.13|
|       24|      3|          506|        41|      19|2022-12-23 21:26:29|   Blue Jeans|   Pants|    173.1|
|        1|      3|          278|         5|       5|2022-12-23 16:41:42|   Red Shorts|  Shorts|    89.75|
|       23|      3|          849|        36|      13|2022-12-23 13:22:55| Green Chinos|   Pants|   150.93|
|        7|      3|          992|        34|       3|2022-12-23 16:47:14|White Sandals|   Shoes|   160.96|
|        7|      3|          703|        19|      13|2022-12-23 22:36:48|White Sandals|   Shoes|   160.96|
|       18|      3|          719|        48|      12|2022-12-23 10:11:29|Black t-shirt|T-Shirts|   102.41|
|       14|      3|          526|        13|       3|2022-12-23 11:57:23|  Red t-shirt|T-Shirts|   121.58|
|        1|      3|          997|        20|      14|2022-12-23 04:02:30|   Red Shorts|  Shorts|    89.75|
|       15|      3|          281|        11|      25|2022-12-23 16:07:45|White t-shirt|T-Shirts|   131.13|
|       23|      3|          691|        48|       2|2022-12-23 08:12:00| Green Chinos|   Pants|   150.93|
|        5|      3|          762|        17|      26|2022-12-23 16:18:27| Black Shorts|  Shorts|    74.58|
|       23|      3|          106|        24|      11|2022-12-23 07:41:50| Green Chinos|   Pants|   150.93|
|        9|      3|           21|        32|       2|2022-12-23 21:15:10|Green Sandals|   Shoes|   137.53|
|       18|      3|          626|        14|      14|2022-12-23 12:55:02|Black t-shirt|T-Shirts|   102.41|
|       15|      3|          219|        11|       5|2022-12-23 13:00:17|White t-shirt|T-Shirts|   131.13|
+---------+-------+-------------+----------+--------+-------------------+-------------+--------+---------+
only showing top 20 rows
```

creating a view and then passing a query to get total sales for store with Id 1

In [6]:
```
merged_df.createOrReplaceTempView("TotalSales")
```

In [7]:
```
output = scSpark.sql("SELECT SUM(Quantity*UnitPrice) from TotalSales WHERE StoreId == 1")
```

In [8]:
```
output.show()
```

```
+--------------------------+
|sum((Quantity * UnitPrice))|
+--------------------------+
|         41264.000000000015|
+--------------------------+
```

Query to get the average amount spent on store with Id 2 by customers

In [9]:
```
output_2 = scSpark.sql("SELECT AVG(Quantity*UnitPrice) from TotalSales WHERE StoreId == 2")
```

In [10]:
```
output_2.show()
```

```
+--------------------------+
|avg((Quantity * UnitPrice))|
+--------------------------+
|          513.4598039215689|
+--------------------------+
```

In [11]:
```
merged_df_2=merged_df.join(df_3,"CustomerId","inner")
```

In [12]:
```
merged_df_2.show()
```

```
+----------+---------+-------+-------------+--------+-------------------+-------------+--------+---------+--------------------+--------------------+
|CustomerId|ProductId|StoreId|TransactionId|Quantity|    TransactionTime|         Name|Category|UnitPrice|                Name|               Email|
+----------+---------+-------+-------------+--------+-------------------+-------------+--------+---------+--------------------+--------------------+
|        35|        3|      3|          454|       3|2022-12-23 17:36:11|  Blue Shorts|  Shorts|   118.88|      Dwayne Johnson|dwayne.johnson@gm...|
|        37|        9|      3|          524|      11|2022-12-23 22:02:51|Green Sandals|   Shoes|   137.53|       Brittany Holt|brittany.holt@exa...|
|         4|        3|      3|          562|       4|2022-12-23 02:51:50|  Blue Shorts|  Shorts|   118.88|       Alevtin Paska|alevtin.paska@exa...|
|        35|       14|      3|          581|      56|2022-12-23 17:05:54|  Red t-shirt|T-Shirts|   121.58|      Dwayne Johnson|dwayne.johnson@gm...|
|        34|       15|      3|          200|      24|2022-12-23 07:15:01|White t-shirt|T-Shirts|   131.13|           Avi Shet|avi.shet@example.com|
|        41|       24|      3|          506|      19|2022-12-23 21:26:29|   Blue Jeans|   Pants|    173.1|         Alice Morin|alice.morin@examp...|
|         5|        1|      3|          278|       5|2022-12-23 16:41:42|   Red Shorts|  Shorts|    89.75|      Charlotte Wong|charlotte.wong@ex...|
|        36|       23|      3|          849|      13|2022-12-23 13:22:55| Green Chinos|   Pants|   150.93|     William Nielsen|william.nielsen@e...|
|        34|        7|      3|          992|       3|2022-12-23 16:47:14|White Sandals|   Shoes|   160.96|           Avi Shet|avi.shet@example.com|
|        19|        7|      3|          703|      13|2022-12-23 22:36:48|White Sandals|   Shoes|   160.96|       Alexia Renaud|alexia.renaud@exa...|
|        48|       18|      3|          719|      12|2022-12-23 10:11:29|Black t-shirt|T-Shirts|   102.41|        Amoli Shenoy|amoli.shenoy@exam...|
|        13|       14|      3|          526|       3|2022-12-23 11:57:23|  Red t-shirt|T-Shirts|   121.58|      Elizabeth Neal|elizabeth.neal@ex...|
|        20|        1|      3|          997|      14|2022-12-23 04:02:30|   Red Shorts|  Shorts|    89.75|         Suzy Gibson|suzy.gibson@examp...|
|        11|       15|      3|          281|      25|2022-12-23 16:07:45|White t-shirt|T-Shirts|   131.13|    Angélique Vennix|angelique.vennix@...|
|        48|       23|      3|          691|       2|2022-12-23 08:12:00| Green Chinos|   Pants|   150.93|        Amoli Shenoy|amoli.shenoy@exam...|
|        17|        5|      3|          762|      26|2022-12-23 16:18:27| Black Shorts|  Shorts|    74.58|Sevastiana Nester...|sevastiana.nester...|
|        24|       23|      3|          106|      11|2022-12-23 07:41:50| Green Chinos|   Pants|   150.93|         Bernd Colin|bernd.colin@examp...|
|        32|        9|      3|           21|       2|2022-12-23 21:15:10|Green Sandals|   Shoes|   137.53|        Ethan Little|ethan.little@exam...|
|        14|       18|      3|          626|      14|2022-12-23 12:55:02|Black t-shirt|T-Shirts|   102.41|      Sylvie Lecomte|sylvie.lecomte@ex...|
|        11|       15|      3|          219|       5|2022-12-23 13:00:17|White t-shirt|T-Shirts|   131.13|    Angélique Vennix|angelique.vennix@...|
+----------+---------+-------+-------------+--------+-------------------+-------------+--------+---------+--------------------+--------------------+
only showing top 20 rows
```

query to get the total amount spent by each customer on each store

In [13]:
```
x=scSpark.sql("SELECT SUM(Quantity*UnitPrice) AS sum,CustomerId From TotalSales GROUP BY CustomerId")
```

In [14]:
```
x.show()
```

```
+------------------+----------+
|               sum|CustomerId|
+------------------+----------+
|           4944.65|         7|
|3237.5400000000004|        15|
|            5317.7|        11|
|           3579.04|        29|
|           5579.95|        34|
|3480.7799999999997|         8|
|          10653.08|        35|
|3269.7800000000007|        16|
|2394.0299999999997|         5|
|           3335.17|        31|
|           8440.65|        17|
|           3824.06|        26|
|1116.8600000000001|         6|
|           2092.48|        19|
|            5060.1|        41|
|           1580.47|        38|
|            522.06|        25|
|3962.9799999999996|        44|
|1895.0700000000002|        48|
|           2768.62|        24|
+------------------+----------+
only showing top 20 rows
```

In [15]:
```
merged_df_3=merged_df_2.join(x,"CustomerId","inner")
merged_df_3.show()
```

```
+----------+---------+-------+-------------+--------+-------------------+-------------+--------+---------+--------------------+--------------------+------------------+
|CustomerId|ProductId|StoreId|TransactionId|Quantity|    TransactionTime|         Name|Category|UnitPrice|                Name|               Email|               sum|
+----------+---------+-------+-------------+--------+-------------------+-------------+--------+---------+--------------------+--------------------+------------------+
|        35|        3|      3|          454|       3|2022-12-23 17:36:11|  Blue Shorts|  Shorts|   118.88|      Dwayne Johnson|dwayne.johnson@gm...|          10653.08|
|        37|        9|      3|          524|      11|2022-12-23 22:02:51|Green Sandals|   Shoes|   137.53|       Brittany Holt|brittany.holt@exa...|3137.8999999999996|
|         4|        3|      3|          562|       4|2022-12-23 02:51:50|  Blue Shorts|  Shorts|   118.88|       Alevtin Paska|alevtin.paska@exa...|           1542.55|
|        35|       14|      3|          581|      56|2022-12-23 17:05:54|  Red t-shirt|T-Shirts|   121.58|      Dwayne Johnson|dwayne.johnson@gm...|          10653.08|
|        34|       15|      3|          200|      24|2022-12-23 07:15:01|White t-shirt|T-Shirts|   131.13|           Avi Shet|avi.shet@example.com|           5579.95|
|        41|       24|      3|          506|      19|2022-12-23 21:26:29|   Blue Jeans|   Pants|    173.1|        Alice Morin|alice.morin@examp...|            5060.1|
|         5|        1|      3|          278|       5|2022-12-23 16:41:42|   Red Shorts|  Shorts|    89.75|     Charlotte Wong|charlotte.wong@ex...|2394.0299999999997|
|        36|       23|      3|          849|      13|2022-12-23 13:22:55| Green Chinos|   Pants|   150.93|    William Nielsen|william.nielsen@e...|           3582.17|
|        34|        7|      3|          992|       3|2022-12-23 16:47:14|White Sandals|   Shoes|   160.96|           Avi Shet|avi.shet@example.com|           5579.95|
|        19|        7|      3|          703|      13|2022-12-23 22:36:48|White Sandals|   Shoes|   160.96|      Alexia Renaud|alexia.renaud@exa...|           2092.48|
|        48|       18|      3|          719|      12|2022-12-23 10:11:29|Black t-shirt|T-Shirts|   102.41|       Amoli Shenoy|amoli.shenoy@exam...|1895.0700000000002|
|        13|       14|      3|          526|       3|2022-12-23 11:57:23|  Red t-shirt|T-Shirts|   121.58|      Elizabeth Neal|elizabeth.neal@ex...|           2585.39|
|        20|        1|      3|          997|      14|2022-12-23 04:02:30|   Red Shorts|  Shorts|    89.75|        Suzy Gibson|suzy.gibson@examp...|            4352.0|
|        11|       15|      3|          281|      25|2022-12-23 16:07:45|White t-shirt|T-Shirts|   131.13|    Angélique Vennix|angelique.vennix@...|            5317.7|
|        48|       23|      3|          691|       2|2022-12-23 08:12:00| Green Chinos|   Pants|   150.93|       Amoli Shenoy|amoli.shenoy@exam...|1895.0700000000002|
|        17|        5|      3|          762|      26|2022-12-23 16:18:27| Black Shorts|  Shorts|    74.58|Sevastiana Nester...|sevastiana.nester...|           8440.65|
|        24|       23|      3|          106|      11|2022-12-23 07:41:50| Green Chinos|   Pants|   150.93|         Bernd Colin|bernd.colin@examp...|           2768.62|
|        32|        9|      3|           21|       2|2022-12-23 21:15:10|Green Sandals|   Shoes|   137.53|        Ethan Little|ethan.little@exam...|           2714.72|
|        14|       18|      3|          626|      14|2022-12-23 12:55:02|Black t-shirt|T-Shirts|   102.41|      Sylvie Lecomte|sylvie.lecomte@ex...|           2329.24|
|        11|       15|      3|          219|       5|2022-12-23 13:00:17|White t-shirt|T-Shirts|   131.13|    Angélique Vennix|angelique.vennix@...|            5317.7|
+----------+---------+-------+-------------+--------+-------------------+-------------+--------+---------+--------------------+--------------------+------------------+
only showing top 20 rows
```

In [16]:
```
merged_df_3.createOrReplaceTempView("Sales")
```

Query to get the customer's email who spent the most

In [17]:
```
output_3=scSpark.sql("SELECT Email FROM Sales WHERE sum =(SELECT MAX(sum) FROM Sales)")
```

In [18]:
```
output_3.limit(1).show()
```

```
+--------------------+
|               Email|
+--------------------+
|dwayne.johnson@gm...|
+--------------------+
```

Query to get the product ID of the top 5 most frequently bought products

In [22]:
```
output_4=scSpark.sql("SELECT ProductId From Sales GROUP BY ProductId ORDER BY SUM(Quantity) desc limit 5")
```

In [23]:
```
output_4.show()
```

```
+---------+
|ProductId|
+---------+
|       14|
|       24|
|       15|
|        5|
|       19|
+---------+
```