Name: - Muhammad Huzaifa Waseem (2303-KHI-DEG-021)

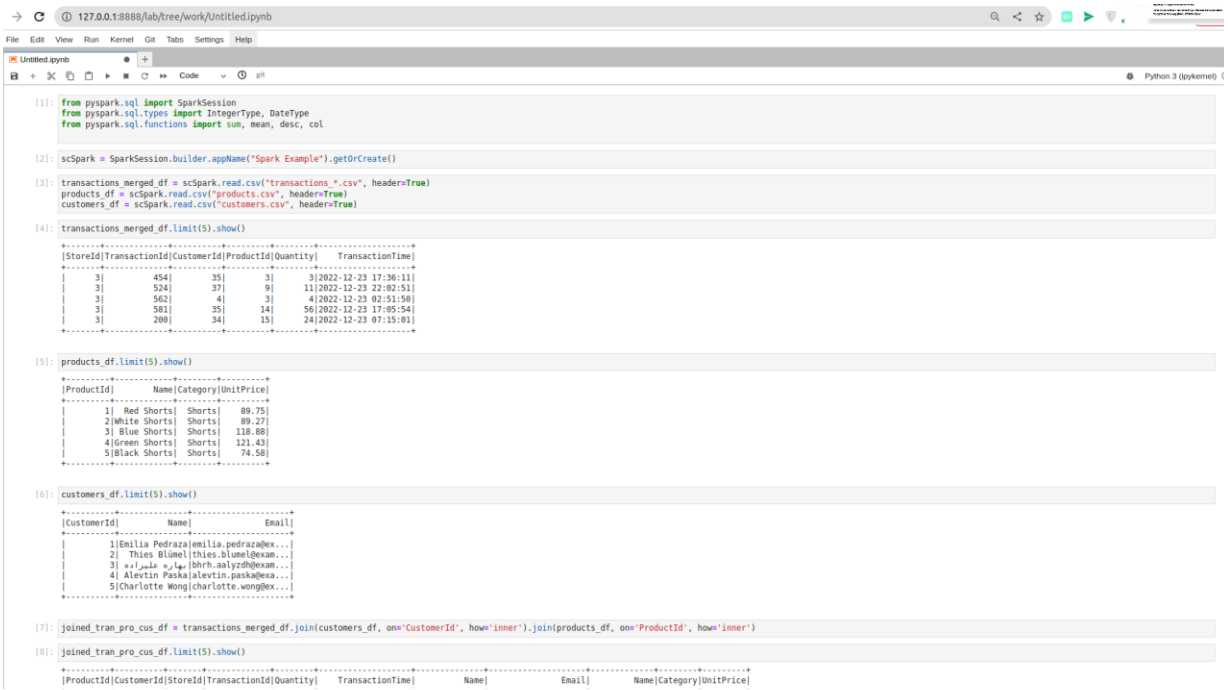
Pair Partner 1: - Muhammad Faizan Rafique (2303.005.KHI.DEG)

Pair Partner 2: - Syed Muhammad Hammad Irshad(2303.KHI.DEG.032)

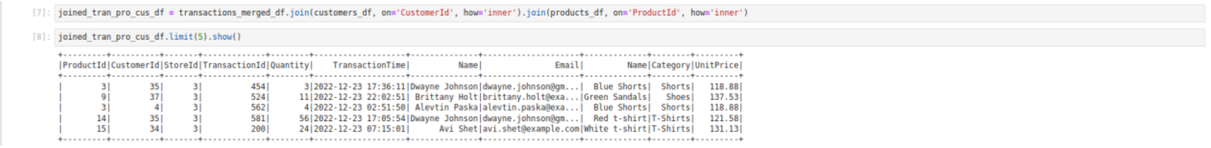
UNIT 5.1:

Assignment

Based on the data contained in *tasks/4_data_pipelines/day_1_introduction/daily_assignment/data* directory, use PySpark to read, filter and join the data from CSV files and answer the following questions:

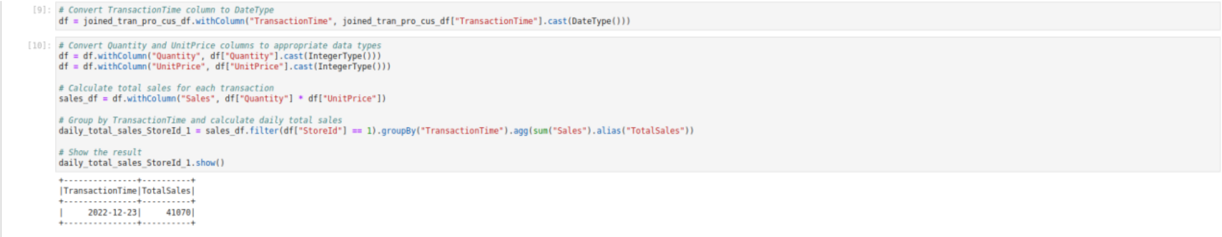

127.0.0.1:8888/lab/tree/work/Untitled.ipynb
File Edit View Run Kernel Git Tabs Settings Help
Untitled.ipynb
Code
Python 3 (ipykernel)

```
[1]: from pyspark.sql import SparkSession
     from pyspark.sql.types import IntegerType, DateType
     from pyspark.sql.functions import sum, mean, desc, col

[2]: scSpark = SparkSession.builder.appName("Spark Example").getOrCreate()

[3]: transactions_merged_df = scSpark.read.csv("transactions_*.csv", header=True)
     products_df = scSpark.read.csv("products.csv", header=True)
     customers_df = scSpark.read.csv("customers.csv", header=True)

[4]: transactions_merged_df.limit(5).show()

     +-------+-------------+----------+---------+--------+-------------------+
     |StoreId|TransactionId|CustomerId|ProductId|Quantity|    TransactionTime|
     +-------+-------------+----------+---------+--------+-------------------+
     |      3|          454|        35|        3|       3|2022-12-23 17:36:11|
     |      3|          524|        37|        9|      11|2022-12-23 22:02:51|
     |      3|          562|         4|        3|       4|2022-12-23 02:51:50|
     |      3|          581|        35|       14|      56|2022-12-23 17:05:54|
     |      3|          200|        34|       15|      24|2022-12-23 07:15:01|
     +-------+-------------+----------+---------+--------+-------------------+

[5]: products_df.limit(5).show()

     +---------+------------+--------+---------+
     |ProductId|        Name|Category|UnitPrice|
     +---------+------------+--------+---------+
     |        1|  Red Shorts|  Shorts|    89.75|
     |        2|White Shorts|  Shorts|    89.27|
     |        3| Blue Shorts|  Shorts|   118.88|
     |        4|Green Shorts|  Shorts|   121.43|
     |        5|Black Shorts|  Shorts|    74.58|
     +---------+------------+--------+---------+

[6]: customers_df.limit(5).show()

     +----------+--------------+--------------------+
     |CustomerId|          Name|               Email|
     +----------+--------------+--------------------+
     |         1|Emilia Pedraza|emilia.pedraza@ex...|
     |         2|  Thies Blümel|thies.blumel@exam...|
     |         3| بهاره علیزاده|bhrh.aalyzdh@exam...|
     |         4| Alevtin Paska|alevtin.paska@exa...|
     |         5|Charlotte Wong|charlotte.wong@ex...|
     +----------+--------------+--------------------+

[7]: joined_tran_pro_cus_df = transactions_merged_df.join(customers_df, on='CustomerId', how='inner').join(products_df, on='ProductId', how='inner')

[8]: joined_tran_pro_cus_df.limit(5).show()

     +---------+----------+-------+-------------+--------+-------------------+--------------+--------------------+------------+--------+---------+
     |ProductId|CustomerId|StoreId|TransactionId|Quantity|    TransactionTime|          Name|               Email|        Name|Category|UnitPrice|
```



```
[7]: joined_tran_pro_cus_df = transactions_merged_df.join(customers_df, on='CustomerId', how='inner').join(products_df, on='ProductId', how='inner')

[8]: joined_tran_pro_cus_df.limit(5).show()

     +---------+----------+-------+-------------+--------+-------------------+--------------+--------------------+------------+--------+---------+
     |ProductId|CustomerId|StoreId|TransactionId|Quantity|    TransactionTime|          Name|               Email|        Name|Category|UnitPrice|
     +---------+----------+-------+-------------+--------+-------------------+--------------+--------------------+------------+--------+---------+
     |        3|        35|      3|          454|       3|2022-12-23 17:36:11|Dwayne Johnson|dwayne.johnson@gm...| Blue Shorts|  Shorts|   118.88|
     |        9|        37|      3|          524|      11|2022-12-23 22:02:51| Brittany Holt|brittany.holt@exa...|Green Sandals|   Shoes|   137.53|
     |        3|         4|      3|          562|       4|2022-12-23 02:51:50| Alevtin Paska|alevtin.paska@exa...| Blue Shorts|  Shorts|   118.88|
     |       14|        35|      3|          581|      56|2022-12-23 17:05:54|Dwayne Johnson|dwayne.johnson@gm...| Red t-shirt|T-Shirts|   121.58|
     |       15|        34|      3|          200|      24|2022-12-23 07:15:01|      Avi Shet|avi.shet@example.com|White t-shirt|T-Shirts|   131.13|
     +---------+----------+-------+-------------+--------+-------------------+--------------+--------------------+------------+--------+---------+
```

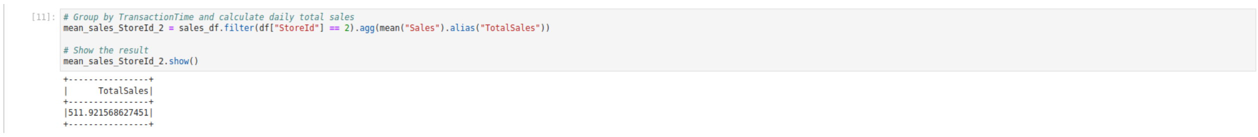
- What are the daily total sales for the store with id 1?



```
[9]:
# Convert TransactionTime column to DateType
df = joined_tran_pro_cus_df.withColumn("TransactionTime", joined_tran_pro_cus_df["TransactionTime"].cast(DateType()))

[10]:
# Convert Quantity and UnitPrice columns to appropriate data types
df = df.withColumn("Quantity", df["Quantity"].cast(IntegerType()))
df = df.withColumn("UnitPrice", df["UnitPrice"].cast(IntegerType()))

# Calculate total sales for each transaction
sales_df = df.withColumn("Sales", df["Quantity"] * df["UnitPrice"])

# Group by TransactionTime and calculate daily total sales
daily_total_sales_StoreId_1 = sales_df.filter(df["StoreId"] == 1).groupBy("TransactionTime").agg(sum("Sales").alias("TotalSales"))

# Show the result
daily_total_sales_StoreId_1.show()
```

```
+---------------+----------+
|TransactionTime|TotalSales|
+---------------+----------+
|     2022-12-23|     41070|
+---------------+----------+
```

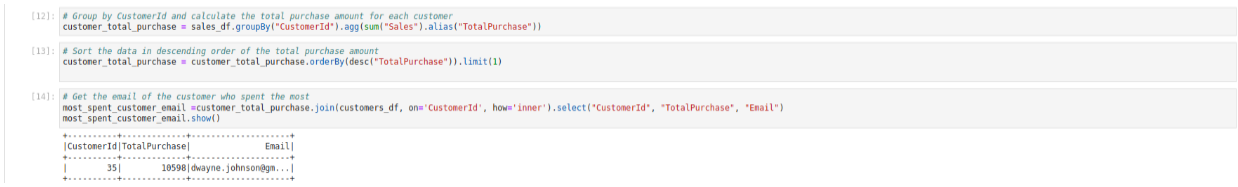
- What are the mean sales for the store with id 2?



```
[11]:
# Group by TransactionTime and calculate daily total sales
mean_sales_StoreId_2 = sales_df.filter(df["StoreId"] == 2).agg(mean("Sales").alias("TotalSales"))

# Show the result
mean_sales_StoreId_2.show()
```

```
+----------------+
|      TotalSales|
+----------------+
|511.921568627451|
+----------------+
```

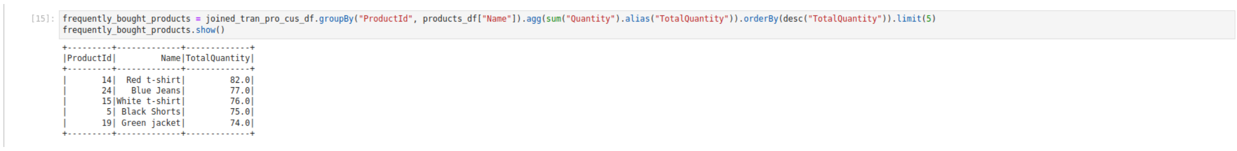
- What is the email of the client who spent the most when summing up purchases from all of the stores?



```
[12]:
# Group by CustomerId and calculate the total purchase amount for each customer
customer_total_purchase = sales_df.groupBy("CustomerId").agg(sum("Sales").alias("TotalPurchase"))

[13]:
# Sort the data in descending order of the total purchase amount
customer_total_purchase = customer_total_purchase.orderBy(desc("TotalPurchase")).limit(1)

[14]:
# Get the email of the customer who spent the most
most_spent_customer_email =customer_total_purchase.join(customers_df, on='CustomerId', how='inner').select("CustomerId", "TotalPurchase", "Email")
most_spent_customer_email.show()
```

```
+----------+-------------+--------------------+
|CustomerId|TotalPurchase|               Email|
+----------+-------------+--------------------+
|        35|        10598|dwayne.johnson@gm...|
+----------+-------------+--------------------+
```

- Which 5 products are most frequently bought across all stores?



```
[15]:
frequently_bought_products = joined_tran_pro_cus_df.groupBy("ProductId", products_df["Name"]).agg(sum("Quantity").alias("TotalQuantity")).orderBy(desc("TotalQuantity")).limit(5)
frequently_bought_products.show()
```

```
+---------+-------------+-------------+
|ProductId|         Name|TotalQuantity|
+---------+-------------+-------------+
|       14|  Red t-shirt|         82.0|
|       24|   Blue Jeans|         77.0|
|       15|White t-shirt|         76.0|
|        5| Black Shorts|         75.0|
|       19| Green jacket|         74.0|
+---------+-------------+-------------+
```