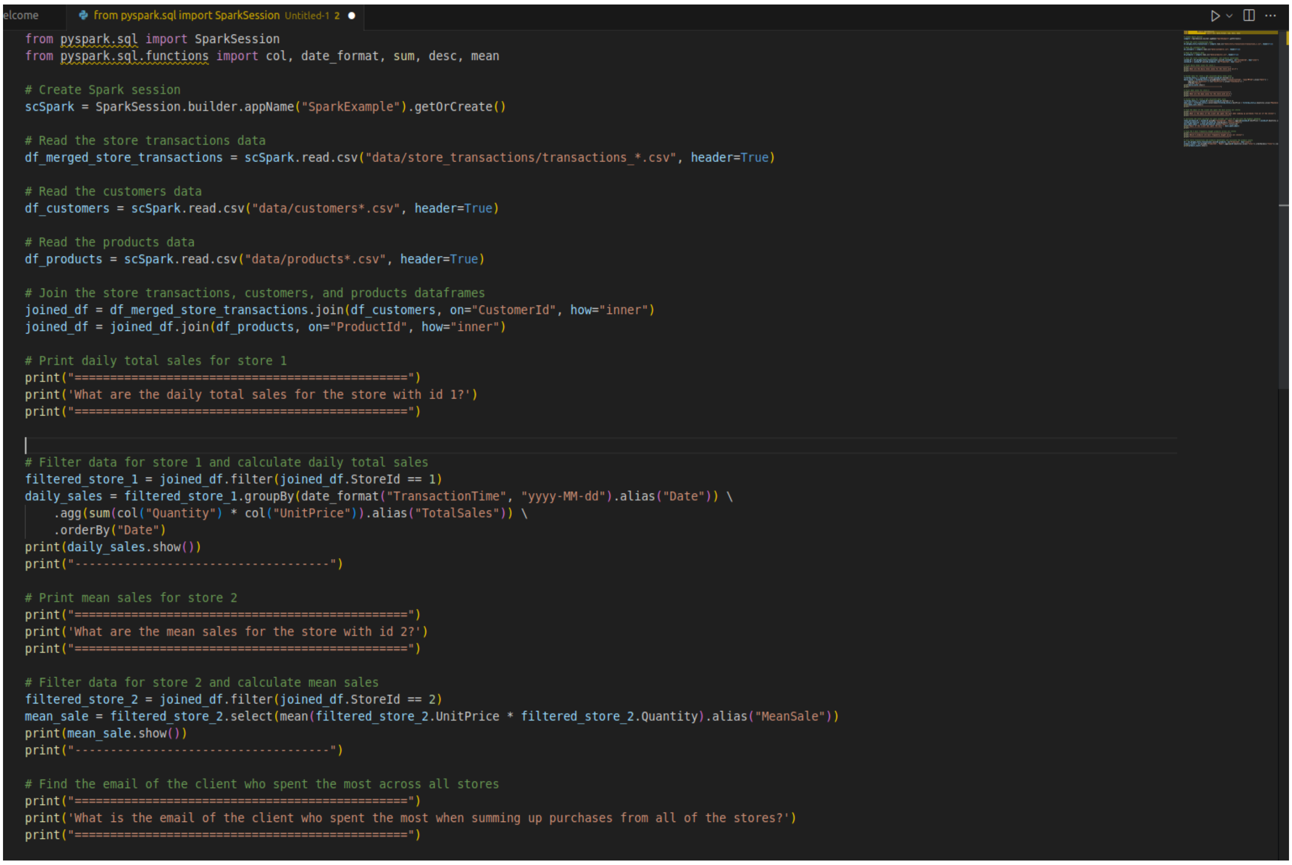
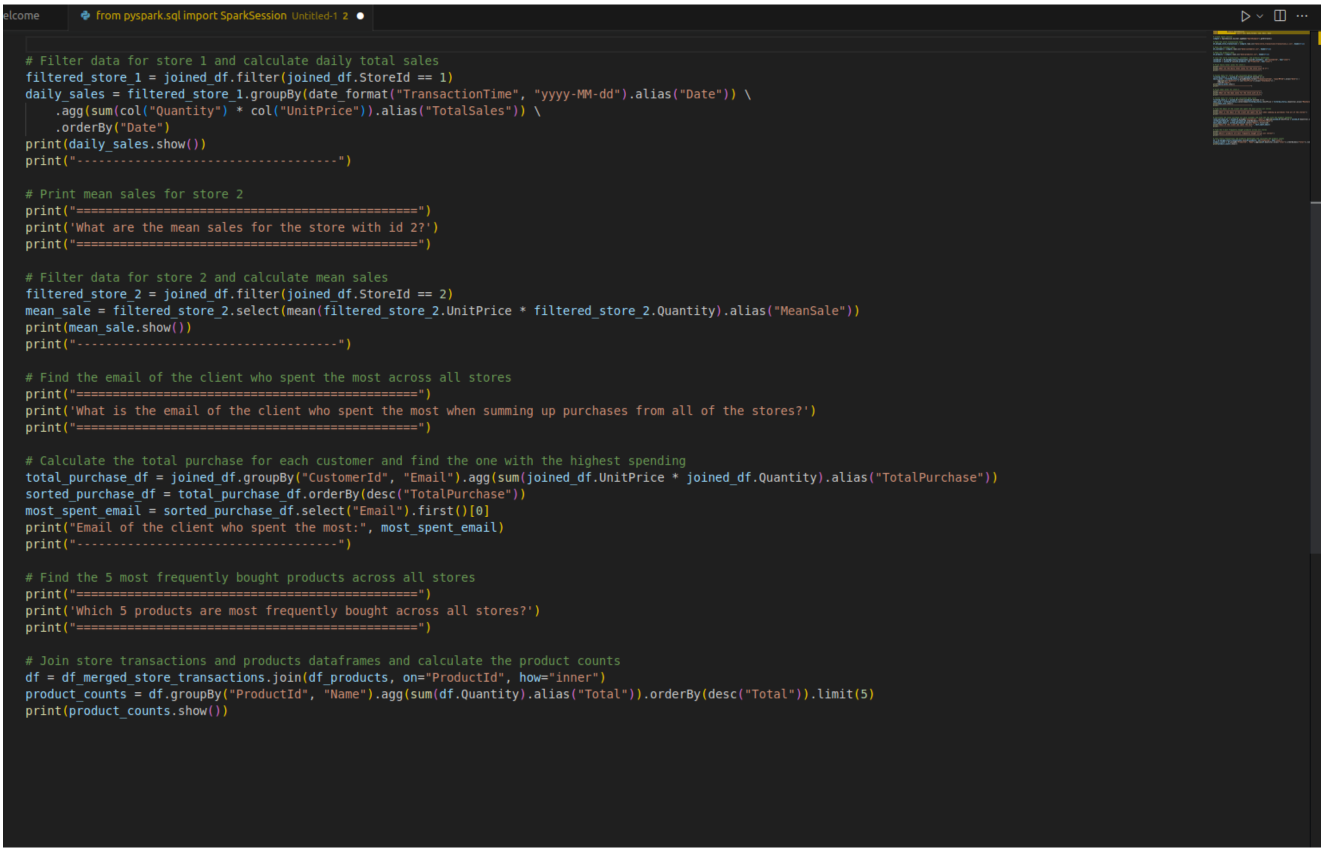
Group Member's Name:-

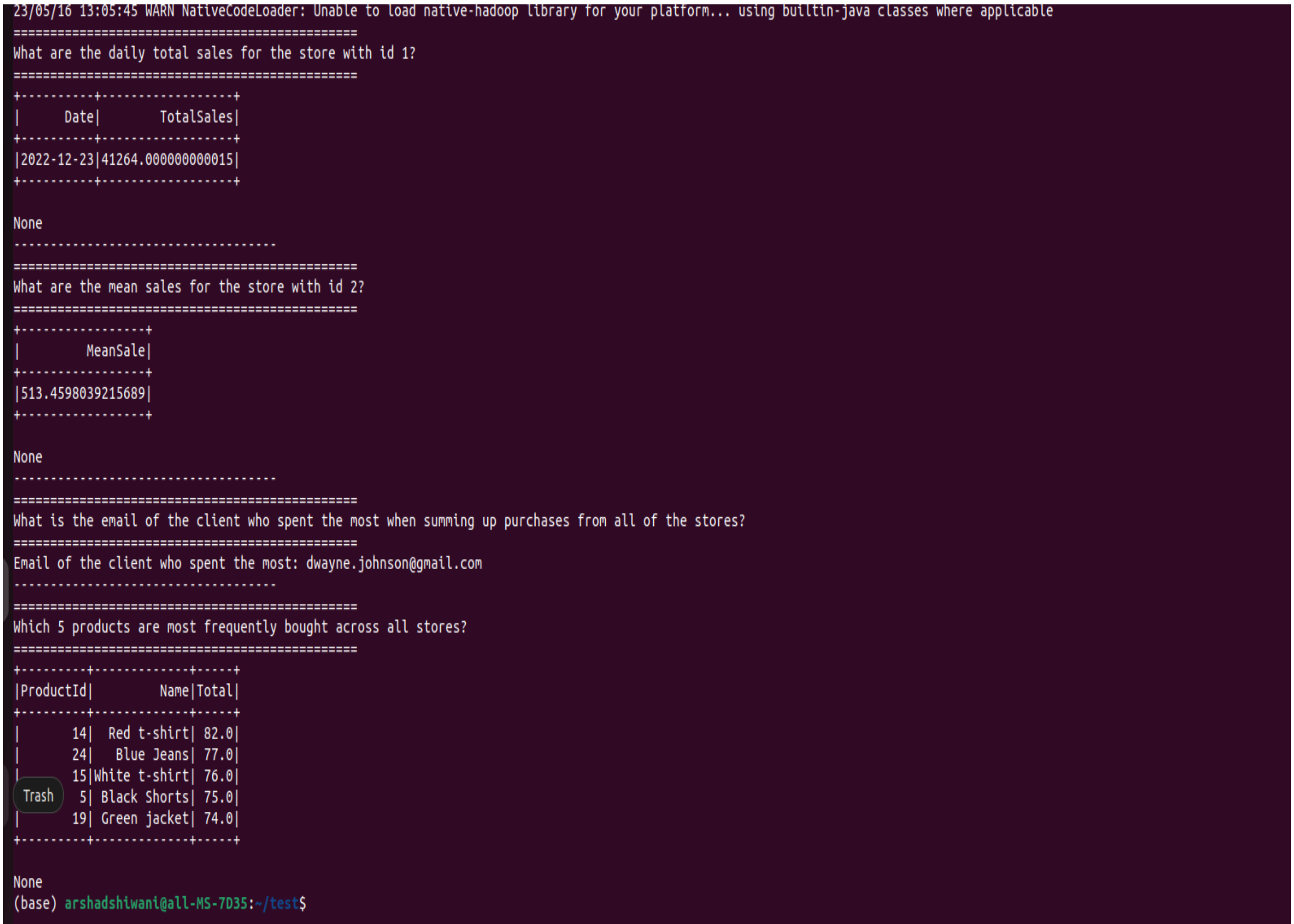
Muhammad Arsalan (2303.KHI.DEG.025)
Abdul Rehman (2303.KHI.DEG.035)
Arshad Shiwani (2303.KHI.DEG.026)



```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, date_format, sum, desc, mean

# Create Spark session
scSpark = SparkSession.builder.appName("SparkExample").getOrCreate()

# Read the store transactions data
df_merged_store_transactions = scSpark.read.csv("data/store_transactions/transactions_*.csv", header=True)

# Read the customers data
df_customers = scSpark.read.csv("data/customers*.csv", header=True)

# Read the products data
df_products = scSpark.read.csv("data/products*.csv", header=True)

# Join the store transactions, customers, and products dataframes
joined_df = df_merged_store_transactions.join(df_customers, on="CustomerId", how="inner")
joined_df = joined_df.join(df_products, on="ProductId", how="inner")

# Print daily total sales for store 1
print("===============================================")
print('What are the daily total sales for the store with id 1?')
print("===============================================")


# Filter data for store 1 and calculate daily total sales
filtered_store_1 = joined_df.filter(joined_df.StoreId == 1)
daily_sales = filtered_store_1.groupBy(date_format("TransactionTime", "yyyy-MM-dd").alias("Date")) \
    .agg(sum(col("Quantity") * col("UnitPrice")).alias("TotalSales")) \
    .orderBy("Date")
print(daily_sales.show())
print("------------------------------------")

# Print mean sales for store 2
print("===============================================")
print('What are the mean sales for the store with id 2?')
print("===============================================")

# Filter data for store 2 and calculate mean sales
filtered_store_2 = joined_df.filter(joined_df.StoreId == 2)
mean_sale = filtered_store_2.select(mean(filtered_store_2.UnitPrice * filtered_store_2.Quantity).alias("MeanSale"))
print(mean_sale.show())
print("------------------------------------")

# Find the email of the client who spent the most across all stores
print("===============================================")
print('What is the email of the client who spent the most when summing up purchases from all of the stores?')
print("===============================================")
```



```python
# Filter data for store 1 and calculate daily total sales
filtered_store_1 = joined_df.filter(joined_df.StoreId == 1)
daily_sales = filtered_store_1.groupBy(date_format("TransactionTime", "yyyy-MM-dd").alias("Date")) \
    .agg(sum(col("Quantity") * col("UnitPrice")).alias("TotalSales")) \
    .orderBy("Date")
print(daily_sales.show())
print("-----------------------------------")

# Print mean sales for store 2
print("=============================================")
print('What are the mean sales for the store with id 2?')
print("=============================================")

# Filter data for store 2 and calculate mean sales
filtered_store_2 = joined_df.filter(joined_df.StoreId == 2)
mean_sale = filtered_store_2.select(mean(filtered_store_2.UnitPrice * filtered_store_2.Quantity).alias("MeanSale"))
print(mean_sale.show())
print("-----------------------------------")

# Find the email of the client who spent the most across all stores
print("=============================================")
print('What is the email of the client who spent the most when summing up purchases from all of the stores?')
print("=============================================")

# Calculate the total purchase for each customer and find the one with the highest spending
total_purchase_df = joined_df.groupBy("CustomerId", "Email").agg(sum(joined_df.UnitPrice * joined_df.Quantity).alias("TotalPurchase"))
sorted_purchase_df = total_purchase_df.orderBy(desc("TotalPurchase"))
most_spent_email = sorted_purchase_df.select("Email").first()[0]
print("Email of the client who spent the most:", most_spent_email)
print("-----------------------------------")

# Find the 5 most frequently bought products across all stores
print("=============================================")
print('Which 5 products are most frequently bought across all stores?')
print("=============================================")

# Join store transactions and products dataframes and calculate the product counts
df = df_merged_store_transactions.join(df_products, on="ProductId", how="inner")
product_counts = df.groupBy("ProductId", "Name").agg(sum(df.Quantity).alias("Total")).orderBy(desc("Total")).limit(5)
print(product_counts.show())
```



```
23/05/16 13:05:45 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
=============================================
What are the daily total sales for the store with id 1?
=============================================
+----------+------------------+
|      Date|        TotalSales|
+----------+------------------+
|2022-12-23|41264.000000000015|
+----------+------------------+

None
-----------------------------------
=============================================
What are the mean sales for the store with id 2?
=============================================
+-----------------+
|         MeanSale|
+-----------------+
|513.4598039215689|
+-----------------+

None
-----------------------------------
=============================================
What is the email of the client who spent the most when summing up purchases from all of the stores?
=============================================
Email of the client who spent the most: dwayne.johnson@gmail.com
-----------------------------------
=============================================
Which 5 products are most frequently bought across all stores?
=============================================
+---------+-------------+-----+
|ProductId|         Name|Total|
+---------+-------------+-----+
|       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|
+---------+-------------+-----+

None
(base) arshadshiwani@all-MS-7D35:~/test$
```