ASSIGNMENT _5.1

Huzaifa Ali(2303.KHI.DEG.016)
Muhammad Safi(2303.KHI.DEG.023)

1)import necessary library and joining data frames

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

```
scSpark = SparkSession.builder.appName("Spark Assignment").getOrCreate()
```

```
transaction_df = scSpark.read.csv("transactions_*.csv", header=True)


customers_df = scSpark.read.csv("customers.csv", header=True)


product_df = scSpark.read.csv("products.csv", header=True)


joined_df = transaction_df.join(customers_df, on='CustomerId', how='inner') \
                          .join(product_df, on='ProductId', how='inner')
```

```
joined_df.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|
|       24|        41|      3|          506|      19|2022-12-23 21:26:29|         Alice Morin|alice.morin@examp...|   Blue Jeans|   Pants|    173.1|
|        1|         5|      3|          278|       5|2022-12-23 16:41:42|      Charlotte Wong|charlotte.wong@ex...|   Red Shorts|  Shorts|    89.75|
|       23|        36|      3|          849|      13|2022-12-23 13:22:55|     William Nielsen|william.nielsen@e...| Green Chinos|   Pants|   150.93|
|        7|        34|      3|          992|       3|2022-12-23 16:47:14|           Avi Shet|avi.shet@example.com|White Sandals|   Shoes|   160.96|
|        7|        19|      3|          703|      13|2022-12-23 22:36:48|       Alexia Renaud|alexia.renaud@exa...|White Sandals|   Shoes|   160.96|
|       18|        48|      3|          719|      12|2022-12-23 10:11:29|       Amoli Shenoy|amoli.shenoy@exam...|Black t-shirt|T-Shirts|   102.41|
|       14|        13|      3|          526|       3|2022-12-23 11:57:23|      Elizabeth Neal|elizabeth.neal@ex...|  Red t-shirt|T-Shirts|   121.58|
|        1|        20|      3|          997|      14|2022-12-23 04:02:30|         Suzy Gibson|suzy.gibson@examp...|   Red Shorts|  Shorts|    89.75|
|       15|        11|      3|          281|      25|2022-12-23 16:07:45|    Angélique Vennix|angelique.vennix@...|White t-shirt|T-Shirts|   131.13|
|       23|        48|      3|          691|       2|2022-12-23 08:12:00|       Amoli Shenoy|amoli.shenoy@exam...| Green Chinos|   Pants|   150.93|
|        5|        17|      3|          762|      26|2022-12-23 16:18:27|Sevastiana Nester...|sevastiana.nester...| Black Shorts|  Shorts|    74.58|
|       23|        24|      3|          106|      11|2022-12-23 07:41:50|        Bernd Colin|bernd.colin@examp...| Green Chinos|   Pants|   150.93|
|        9|        32|      3|           21|       2|2022-12-23 21:15:10|       Ethan Little|ethan.little@exam...|Green Sandals|   Shoes|   137.53|
|       18|        14|      3|          626|      14|2022-12-23 12:55:02|     Sylvie Lecomte|sylvie.lecomte@ex...|Black t-shirt|T-Shirts|   102.41|
|       15|        11|      3|          219|       5|2022-12-23 13:00:17|    Angélique Vennix|angelique.vennix@...|White t-shirt|T-Shirts|   131.13|
+---------+----------+-------+-------------+--------+-------------------+--------------------+--------------------+-------------+--------+---------+
only showing top 20 rows
```

2) Casting datatype

```
joined_df = joined_df.withColumn("TransactionTime", joined_df["TransactionTime"].cast(DateType()))
```

```
joined_df = joined_df.withColumn(
    "Quantity", joined_df["Quantity"].cast(IntegerType())
)
```

```
joined_df = joined_df.withColumn(
    "UnitPrice", joined_df["UnitPrice"].cast(IntegerType())
)
```

3) The daily total sales for the store with id 1

```
store_1_Sales_df = joined_df.filter(col('StoreId') == 1).groupBy('TransactionTime').agg(sum(joined_df.Quantity * joined_df.UnitPrice))
```

```
store_1_Sales_df.show()
```

```
+---------------+--------------------------+
|TransactionTime|sum((Quantity * UnitPrice))|
+---------------+--------------------------+
|     2022-12-23|                     41070|
+---------------+--------------------------+
```

4) The mean sales for the store with id 2

```
store_2_mean_sales = joined_df.filter(col('StoreId') == 2).agg(mean(joined_df.Quantity * joined_df.UnitPrice))
```

```
store_2_mean_sales.show()
```

```
+--------------------------+
|avg((Quantity * UnitPrice))|
+--------------------------+
|          511.921568627451|
+--------------------------+
```

5) the email of the client who spent the most when summing up purchases from all of the stores

```
total_spent_df = joined_df.groupBy('CustomerId').agg(sum(joined_df.Quantity * joined_df.UnitPrice).alias('TotalSpent'))

customer_max_spent = total_spent_df.orderBy(desc('TotalSpent')).first()

email_of_max_spent_customer = customers_df.filter(col('CustomerId') == customer_max_spent['CustomerId']).select('Email')

email_of_max_spent_customer.show()
```

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

6) 5 products are most frequently bought across all stores

```
product_counts_df = joined_df.groupBy('ProductId').agg(sum('Quantity'))

sorted_product_counts_df = product_counts_df.orderBy(desc('sum(Quantity)'))

top_products =sorted_product_counts_df.limit(5)

top_products_with_details = top_products.join(product_df, on='ProductId', how='inner')

result = top_products_with_details.select('ProductId', 'Name','sum(Quantity)').orderBy(desc('sum(Quantity)'))

result.show()
```

```
+---------+-------------+-------------+
|ProductId|         Name|sum(Quantity)|
+---------+-------------+-------------+
|       14|  Red t-shirt|           82|
|       24|   Blue Jeans|           77|
|       15|White t-shirt|           76|
|        5| Black Shorts|           75|
|       19| Green jacket|           74|
+---------+-------------+-------------+
```