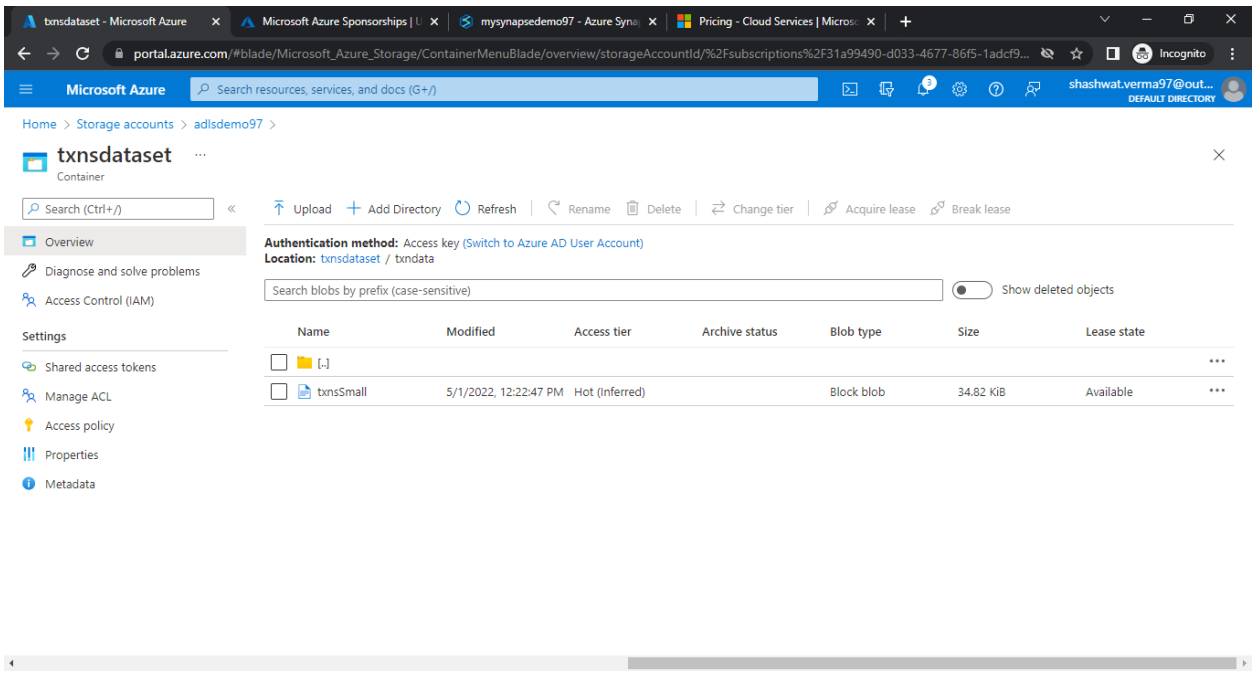
Concept Assignment 1 : Azure Data Engineering

Perform the following using txns Dataset using

1. **SQLPool Internal Table**
2. **SparkPool**
3. **PySpark Notebook in Databricks**
4. **Table Method in Databricks**

a. Find the total revenue generated based on category
b. Find the total number of transactions done by card
c. Find the highest selling category
d. Find the lowest selling category

In the 'adlsdemo97' storage account we created a container named 'txnsdataset'. Inside txns data set we created a directory 'txndata' where we uploaded 'txnsSmall' file as shown in screenshot below:



1. SQLPool Internal Table

-- Created a seperate user for load operations

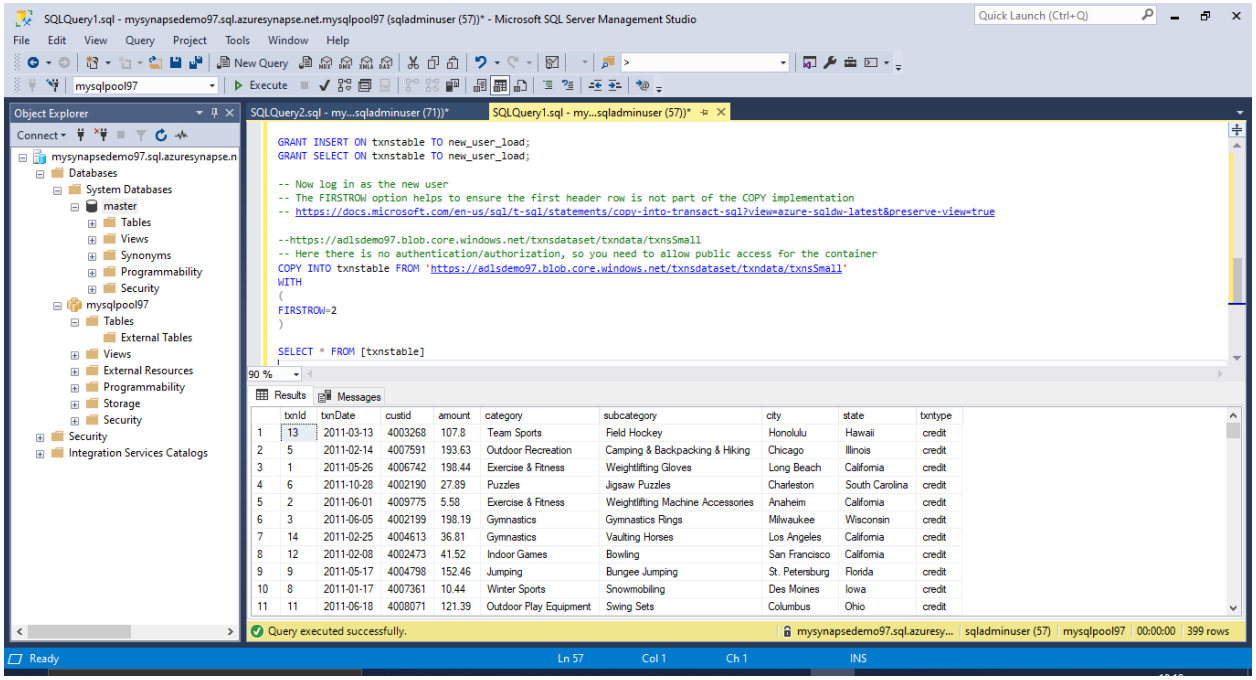
-- This has to be run in the master database

```
CREATE LOGIN new_user_load WITH PASSWORD = 'Azure@123';

-- SQLPool database
CREATE USER new_user_load FOR LOGIN new_user_load;
GRANT ADMINISTER DATABASE BULK OPERATIONS TO new_user_load;
GRANT CREATE TABLE TO new_user_load;
GRANT ALTER ON SCHEMA::dbo TO new_user_load;

CREATE WORKLOAD GROUP DataLoads
WITH (
   MIN_PERCENTAGE_RESOURCE = 100
   ,CAP_PERCENTAGE_RESOURCE = 100
   ,REQUEST_MIN_RESOURCE_GRANT_PERCENT = 100
   );

CREATE WORKLOAD CLASSIFIER [ELTLogin]
WITH (
       WORKLOAD_GROUP = 'DataLoads'
   ,MEMBERNAME = 'new_user_load'
);
-- Create a normal table
-- Login as the new user and create the table
-- Here I have added more constraints when it comes to the width of the data type

CREATE TABLE [txnstable]
(
   [txnId] [bigint] NULL,
         [txnDate] [date] NULL,
         [custid] [bigint] NULL,
         [amount] [float] NULL,
         [category] [varchar](100) NULL,
         [subcategory] [varchar](200) NULL,
         [city] [varchar](100) NULL,
         [state] [varchar](100) NULL,
         [txntype] [varchar](20) NULL)

-- Grant the required privileges to the new user

GRANT INSERT ON txnstable TO new_user_load;
GRANT SELECT ON txnstable TO new_user_load;
```

-- Now log in as the new user
-- The FIRSTROW option helps to ensure the first header row is not part of the COPY implementation
--
https://docs.microsoft.com/en-us/sql/t-sql/statements/copy-into-transact-sql?view=azure-sqldw-latest&preserve-view=true

--https://adlsdemo97.blob.core.windows.net/txnsdataset/txndata/txnsSmall
-- Here there is no authentication/authorization, so we need to allow public access for the container
COPY INTO txnstable FROM
'https://adlsdemo97.blob.core.windows.net/txnsdataset/txndata/txnsSmall'
WITH
(
FIRSTROW=2)

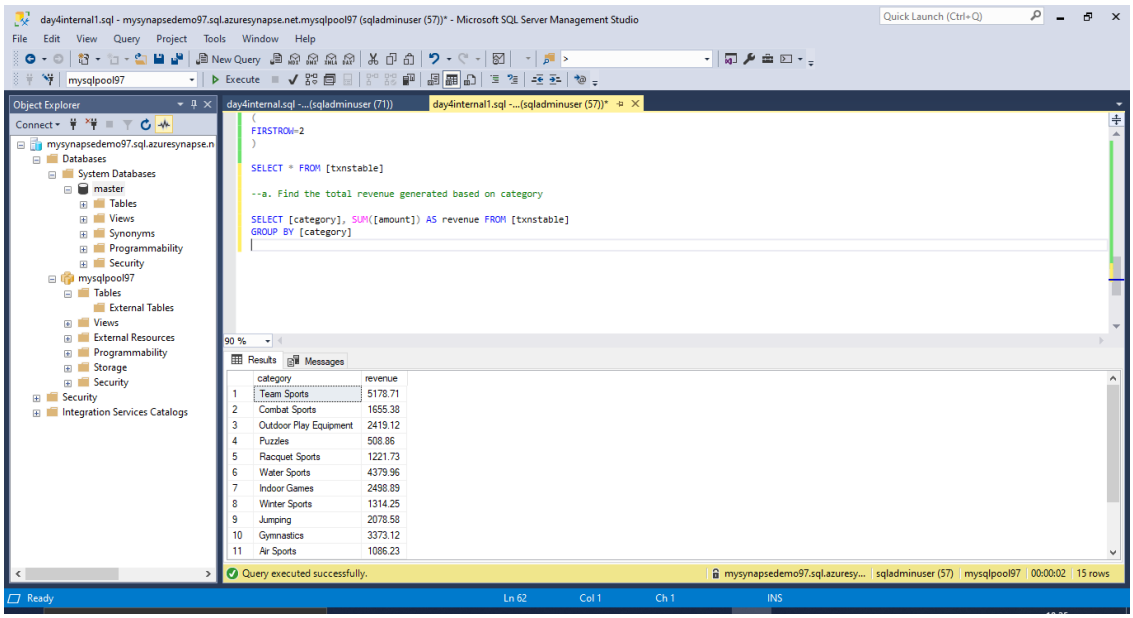
SELECT * FROM [txnstable]



Problems:

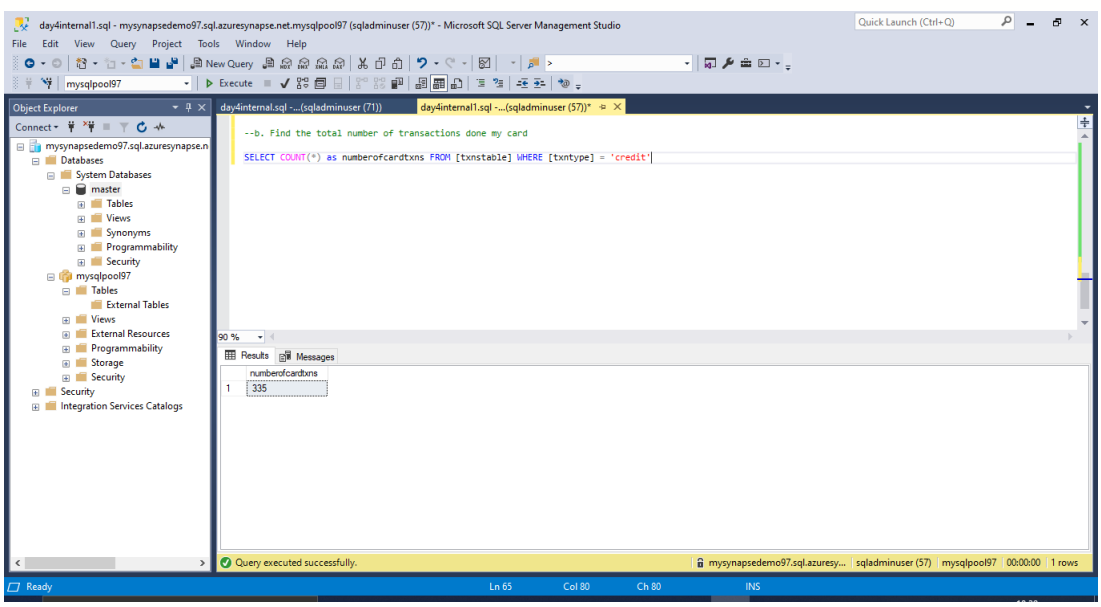
--a. Find the total revenue generated based on category

SELECT [category], SUM([amount]) AS revenue FROM [txnstable]
GROUP BY [category]



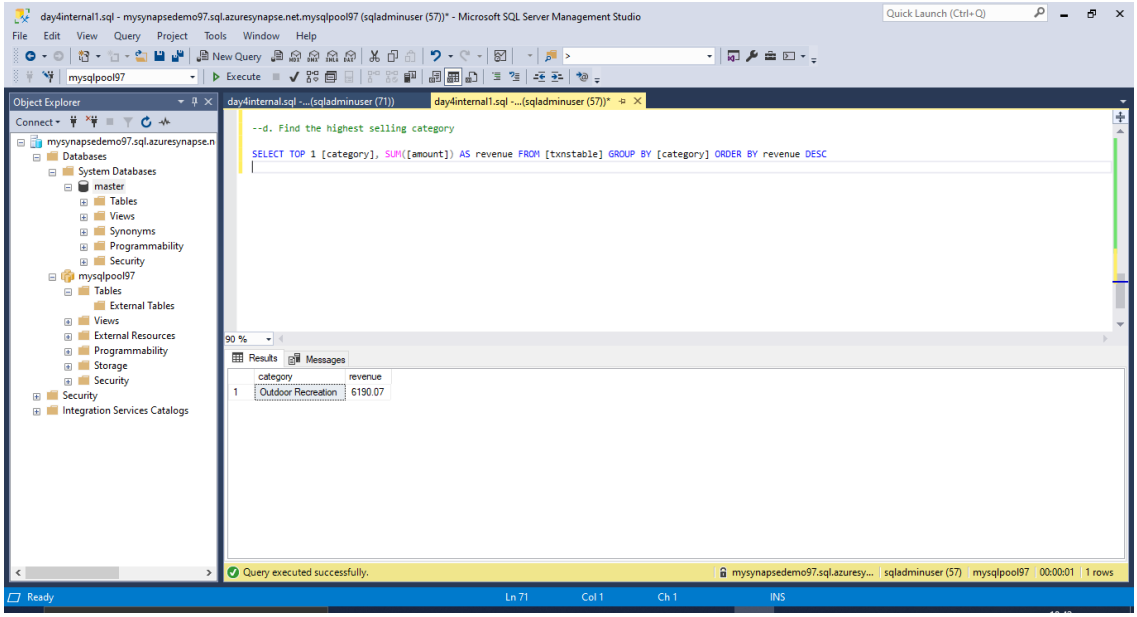
--b. Find the total number of transactions done my card

SELECT COUNT(*) as numberofcardtxns FROM [txnstable] WHERE [txntype] = 'credit'



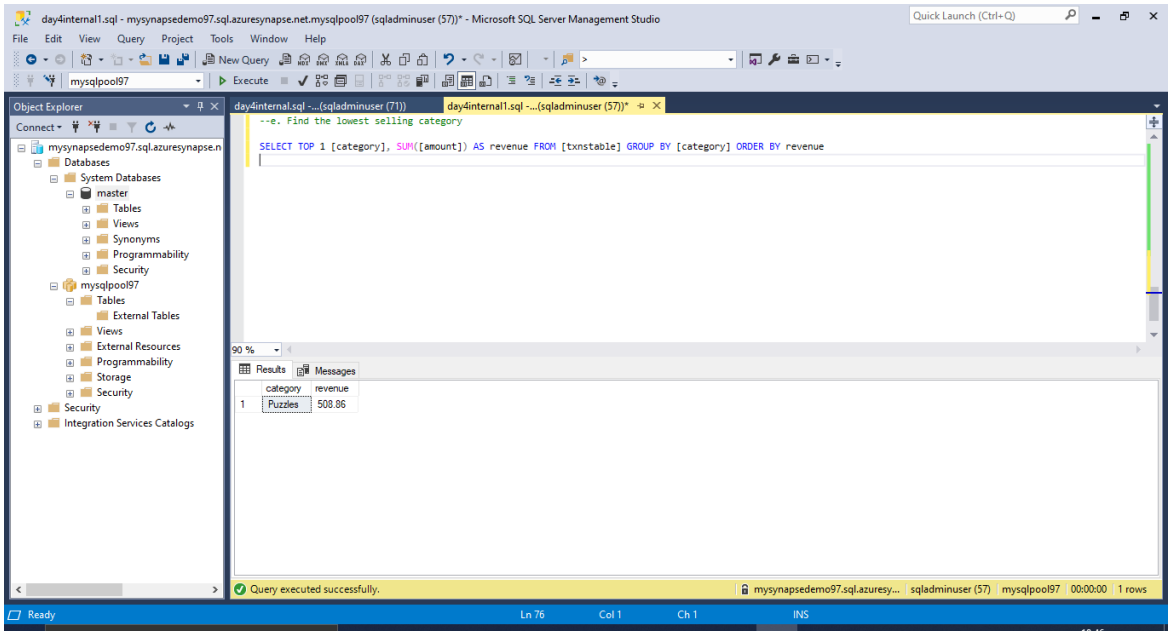
--c. Find the highest selling category

SELECT TOP 1 [category], SUM([amount]) AS revenue FROM [txnstable] GROUP BY [category] ORDER BY revenue DESC



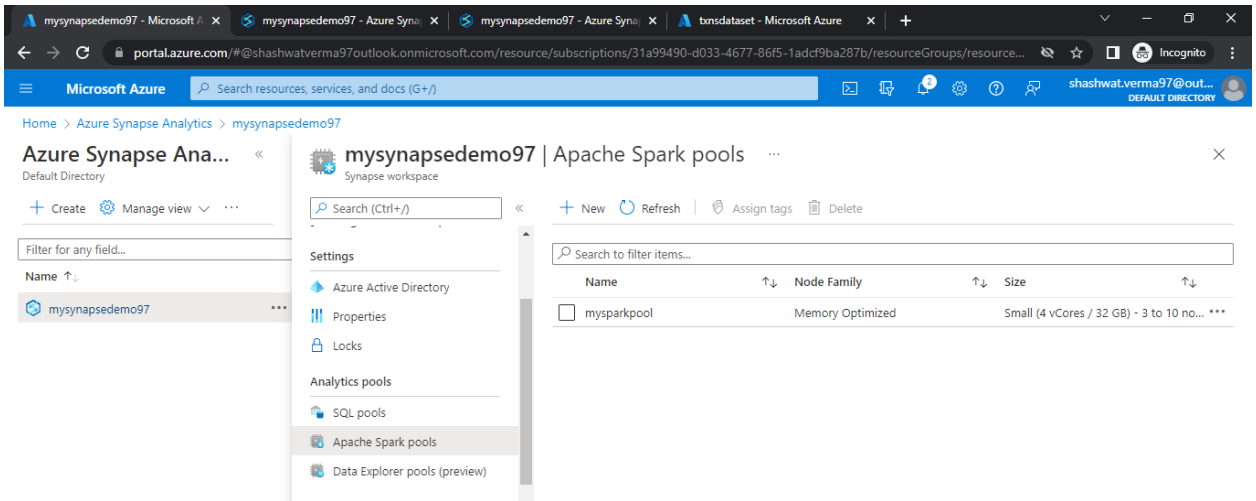
--d. Find the lowest selling category

SELECT TOP 1 [category], SUM([amount]) AS revenue FROM [txnstable] GROUP BY [category] ORDER BY revenue



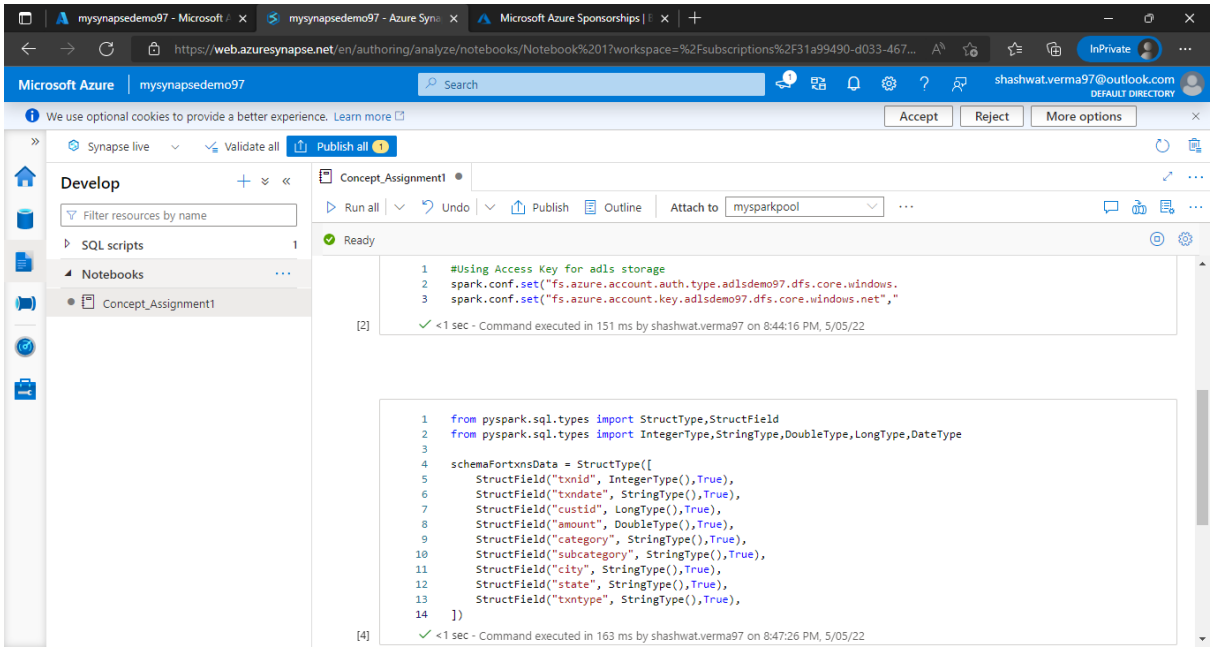
2. SparkPool

Apache spark pool named 'mysparkpool' was created 'mysynapsedemo97' workspace

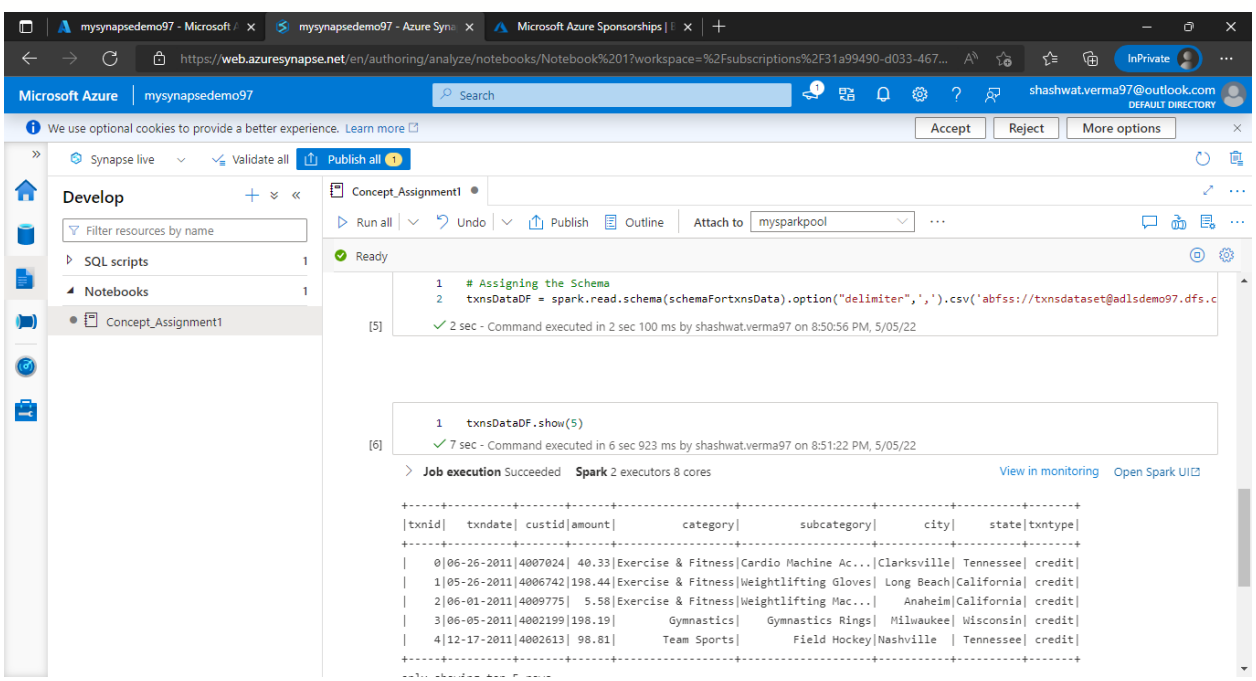


Source Code:

```
#Using Access Key for adls storage
spark.conf.set("fs.azure.account.auth.type.adlsdemo97.dfs.core.windows.net","SharedKey")
spark.conf.set("fs.azure.account.key.adlsdemo97.dfs.core.windows.net","IgWwisxE9bIfEgv5v521p5nE9gb2dX8rl9ZsFDiJAXRv4Y2n+IXderAkHeMJHYLbc1hOFiRPcwZT+AStzQuYeA==")

# Create a Schema Programatically and Assign the same during DataFrame Creation
from pyspark.sql.types import StructType,StructField
from pyspark.sql.types import IntegerType,StringType,DoubleType,LongType,DateType

schemaFortxnsData = StructType([
    StructField("txnid", IntegerType(),True),
    StructField("txndate", StringType(),True),
    StructField("custid", LongType(),True),
    StructField("amount", DoubleType(),True),
    StructField("category", StringType(),True),
    StructField("subcategory", StringType(),True),
    StructField("city", StringType(),True),
    StructField("state", StringType(),True),
    StructField("txntype", StringType(),True),
])
```




```
# Assigning the Schema
txnsDataDF =
spark.read.schema(schemaFortxnsData).option("delimiter",',').csv('abf
ss://txnsdataset@adlsdemo97.dfs.core.windows.net/txndata/txnsSmall')
txnsDataDF.show(5)
```

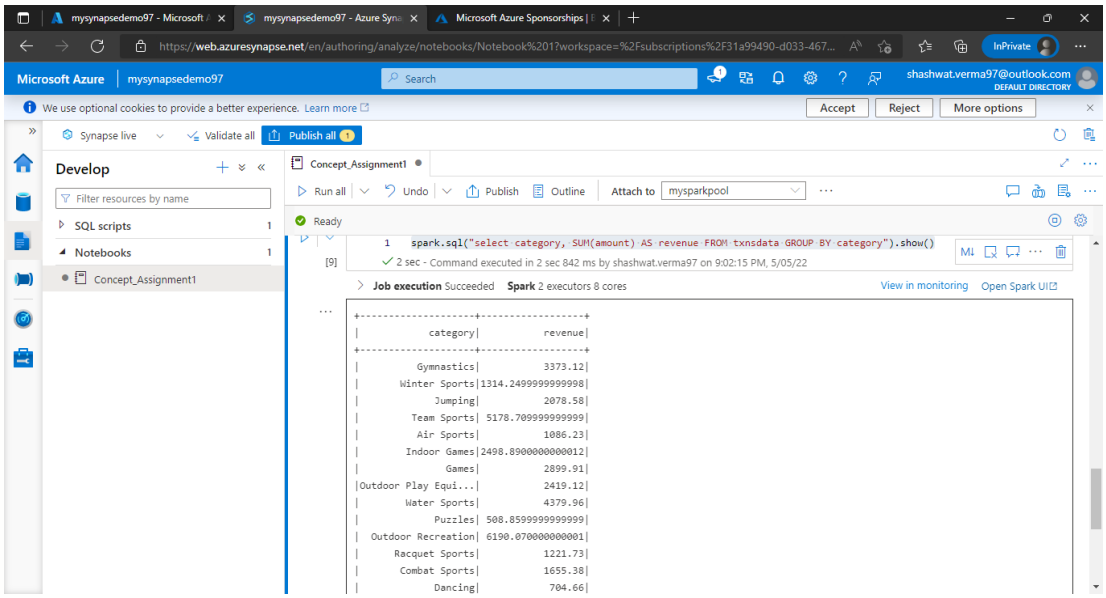



```
# Register this DataFrame as a Temporary Table in Apache Spark
txnsDataDF.registerTempTable("txnsdata")
```

Problems:

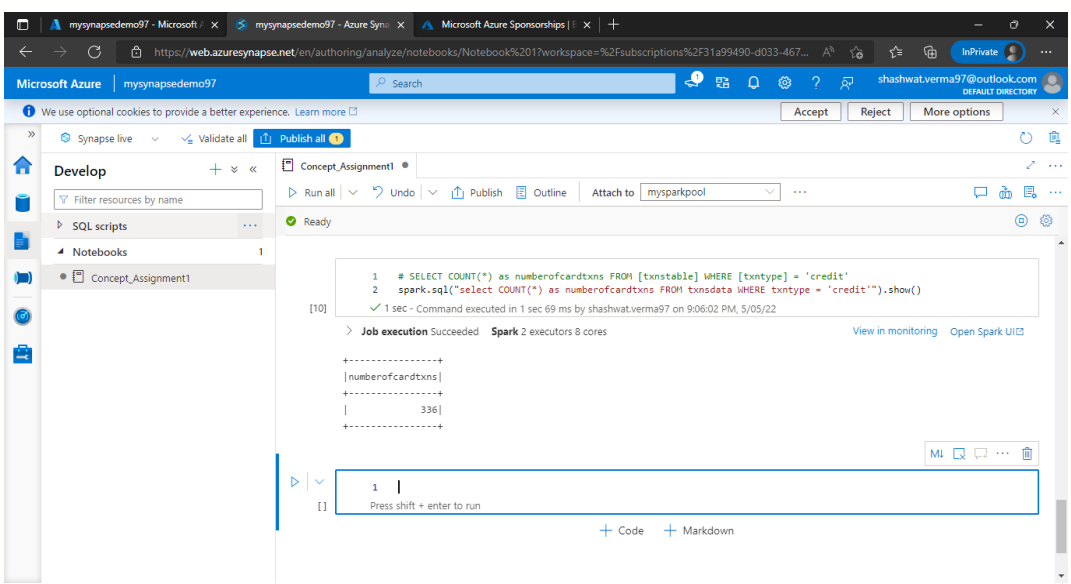
a. Find the total revenue generated based on category

```
spark.sql("select category, SUM(amount) AS revenue FROM txnsdata GROUP BY category").show()
```



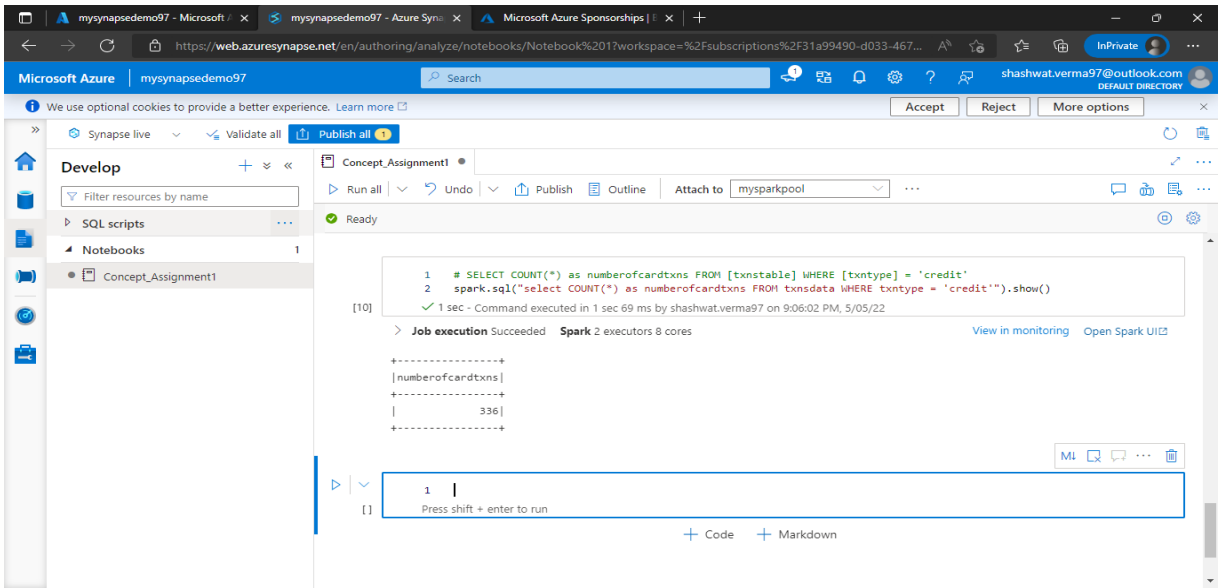
b. Find the total number of transactions done my card

```
spark.sql("select COUNT(*) as numberofcardtxns FROM txnsdata WHERE txntype = 'credit'").show()
```



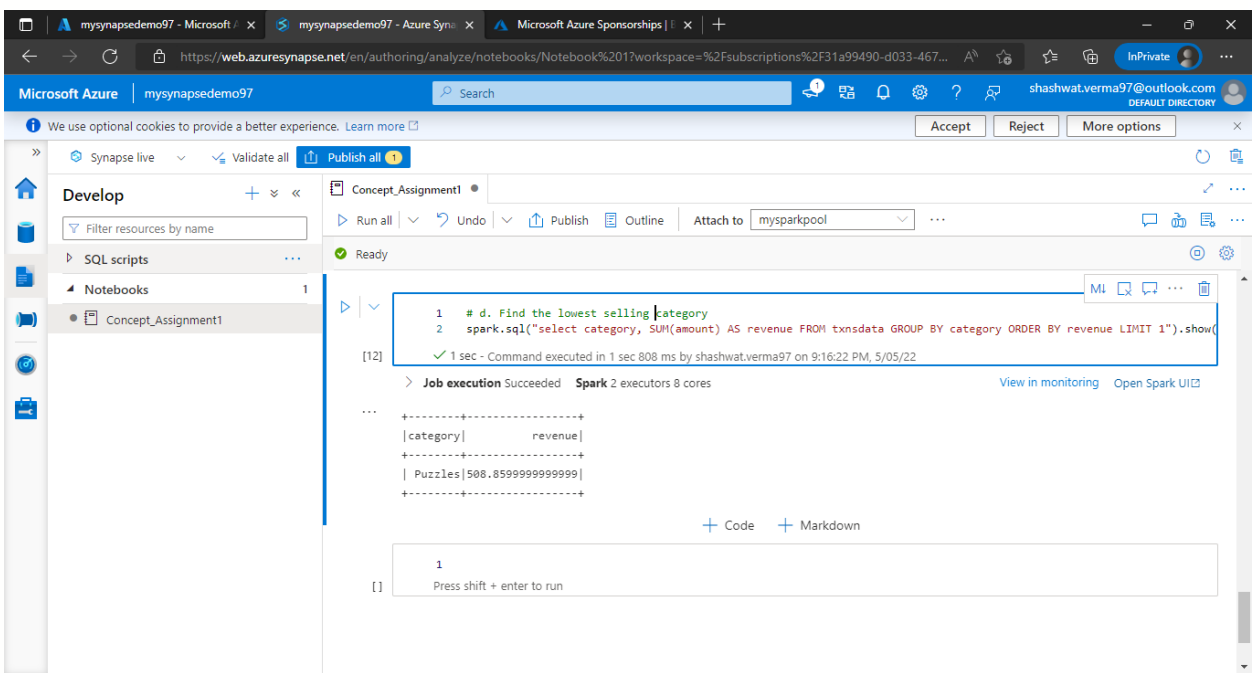
c. Find the highest selling category

```
spark.sql("select category, SUM(amount) AS revenue FROM txnsdata
GROUP BY category ORDER BY revenue DESC LIMIT 1").show()
```



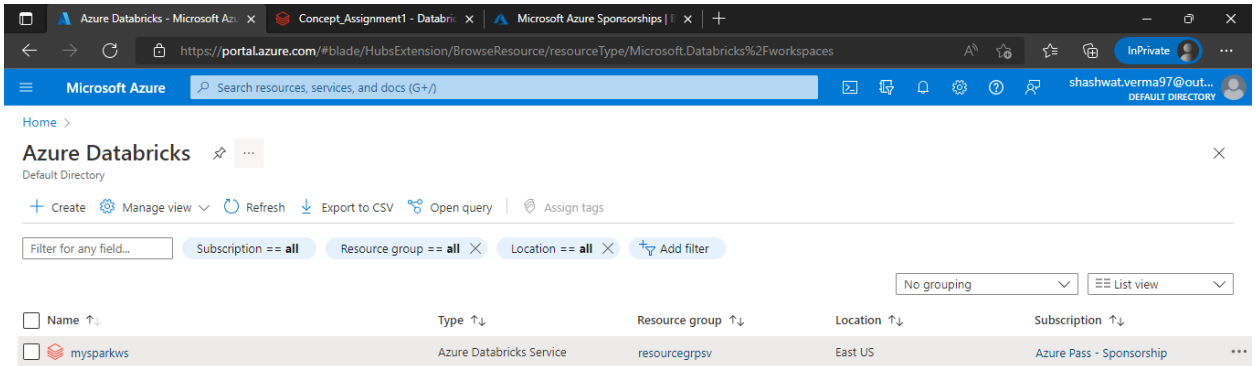
d. Find the lowest selling category

```
spark.sql("select category, SUM(amount) AS revenue FROM txnsdata
GROUP BY category ORDER BY revenue LIMIT 1").show()
```



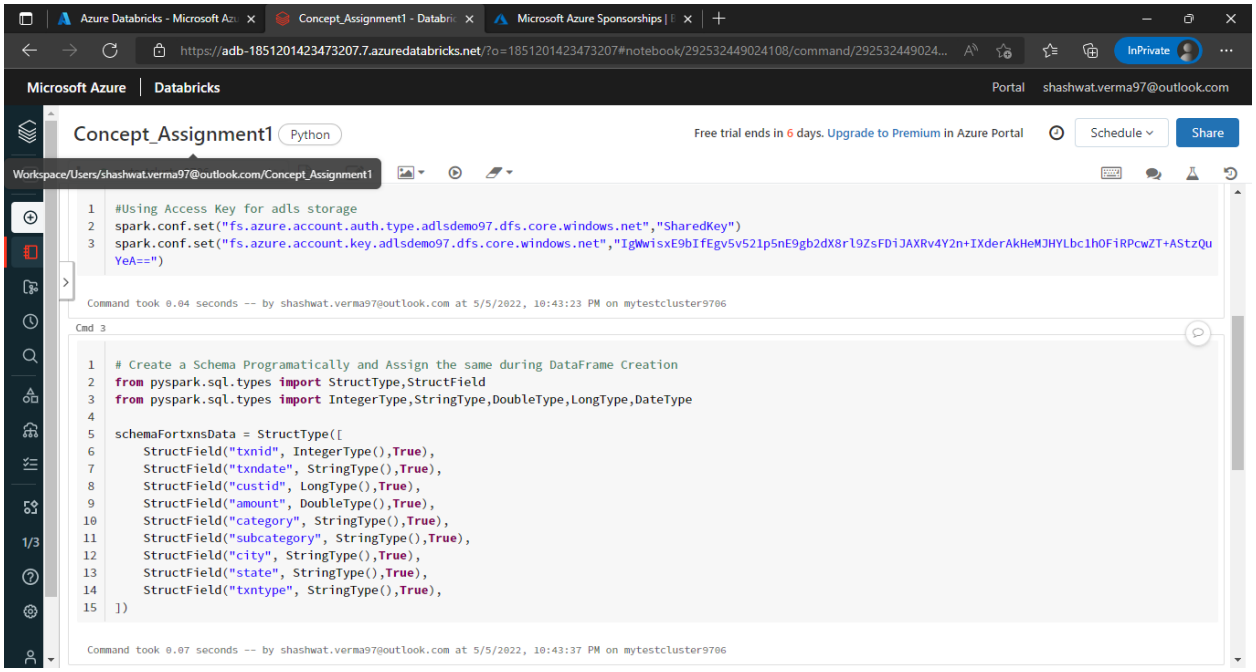
3. PySpark Notebook in Databricks

We create a Azure Databricks workspace 'mysparkws' which is launched and created a cluster 'mytestcluster9706' inside which we created a notebook 'Concept_Assignment1'

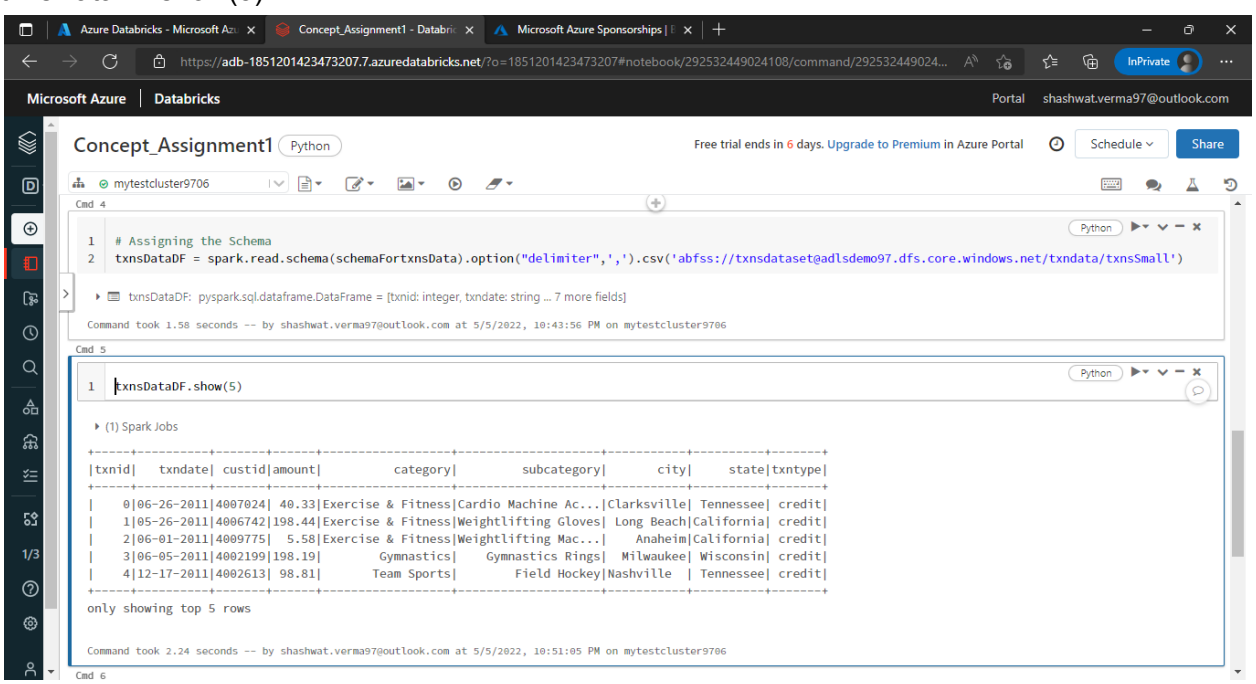


Source Code:

```
#Using Access Key for adls storage
spark.conf.set("fs.azure.account.auth.type.adlsdemo97.dfs.core.windows.net","SharedKey")
spark.conf.set("fs.azure.account.key.adlsdemo97.dfs.core.windows.net","IgWwisxE9bIfEgv5v52
1p5nE9gb2dX8rl9ZsFDiJAXRv4Y2n+IXderAkHeMJHYLbc1hOFiRPcwZT+AStzQuYeA==")

# Create a Schema Programatically and Assign the same during DataFrame Creation
from pyspark.sql.types import StructType,StructField
from pyspark.sql.types import IntegerType,StringType,DoubleType,LongType,DateType

schemaFortxnsData = StructType([
    StructField("txnid", IntegerType(),True),
    StructField("txndate", StringType(),True),
    StructField("custid", LongType(),True),
    StructField("amount", DoubleType(),True),
    StructField("category", StringType(),True),
    StructField("subcategory", StringType(),True),
    StructField("city", StringType(),True),
    StructField("state", StringType(),True),
    StructField("txntype", StringType(),True),
])
# Assigning the Schema
txnsDataDF =
spark.read.schema(schemaFortxnsData).option("delimiter",',').csv('abfss://txnsdataset@adlsde
mo97.dfs.core.windows.net/txndata/txnsSmall')
```



# Assigning the Schema
txnsDataDF = spark.read.schema(schemaFortxnsData).option("delimiter",',').csv('abfss://txnsdataset@adlsdemo97.dfs.core.windows.net/txndata/txnsSmall')
txnsDataDF.show(5)

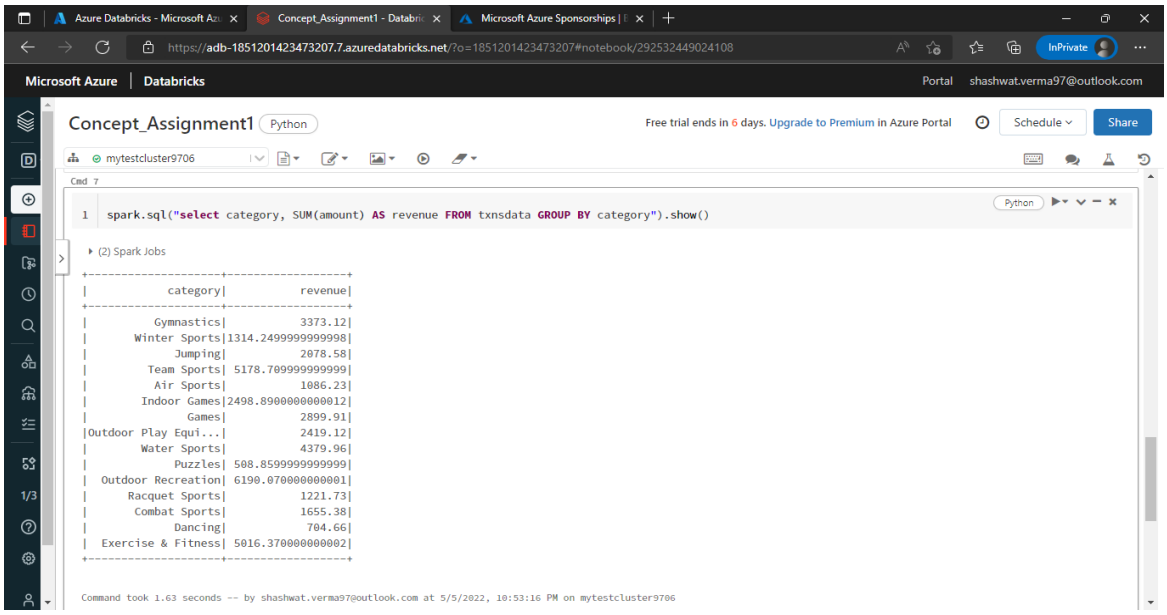


# Register this DataFrame as a Temporary Table in Apache Spark
txnsDataDF.registerTempTable("txnsdata")

Problems:

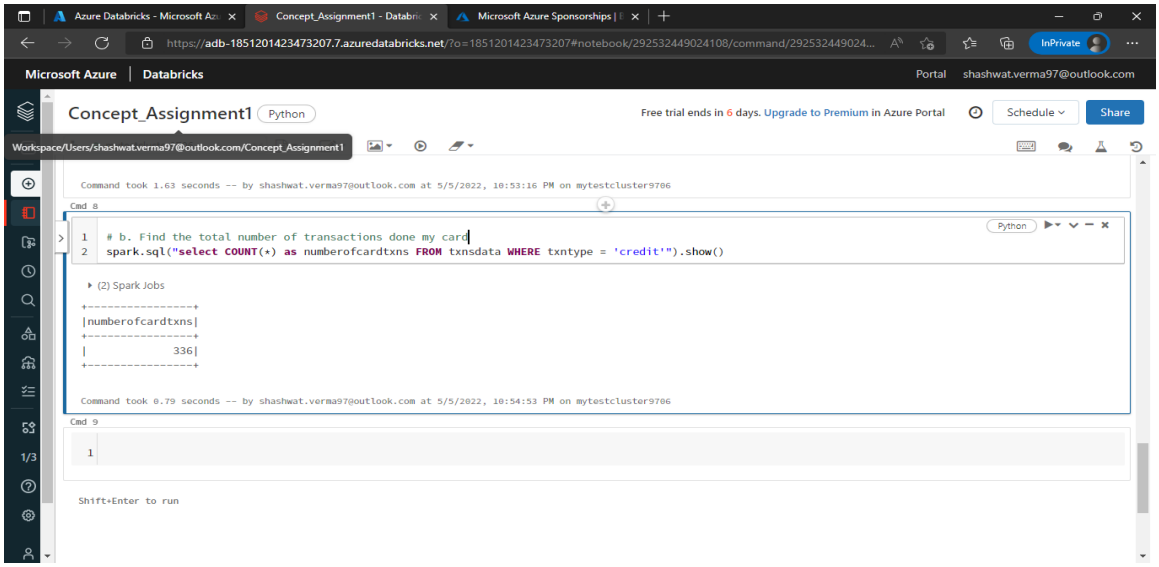
a. Find the total revenue generated based on category

spark.sql("select category, SUM(amount) AS revenue FROM txnsdata GROUP BY category").show()



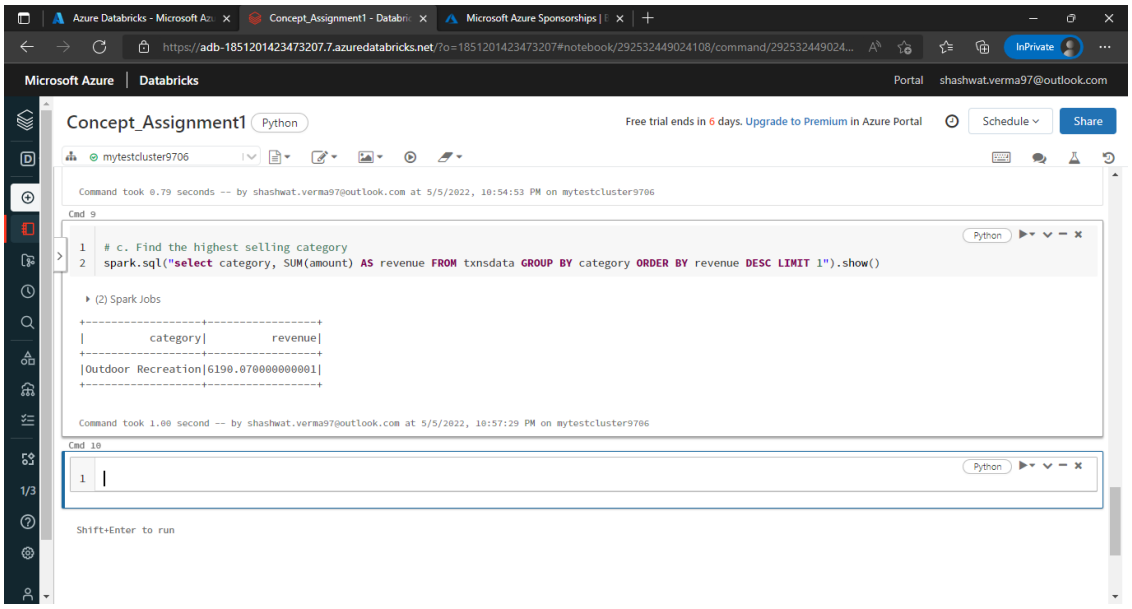
b. Find the total number of transactions done my card

spark.sql("select COUNT(*) as numberofcardtxns FROM txnsdata WHERE txntype = 'credit'").show()



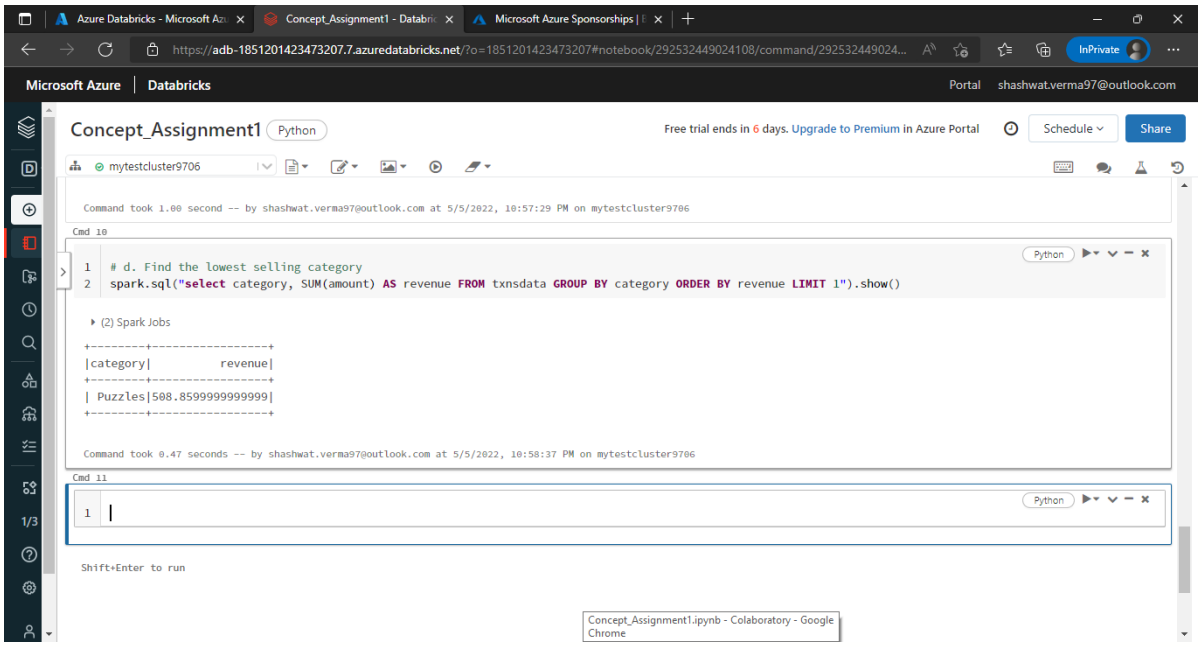
c. Find the highest selling category

spark.sql("select category, SUM(amount) AS revenue FROM txnsdata GROUP BY category ORDER BY revenue DESC LIMIT 1").show()



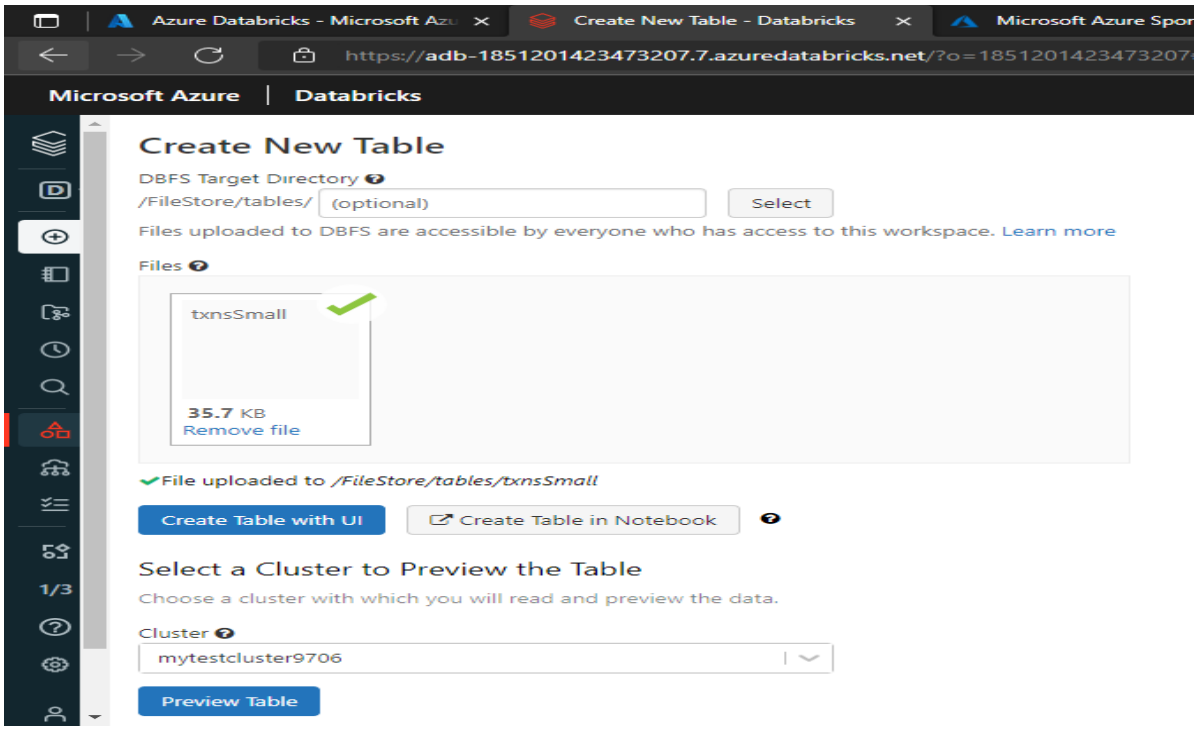
d. Find the lowest selling category

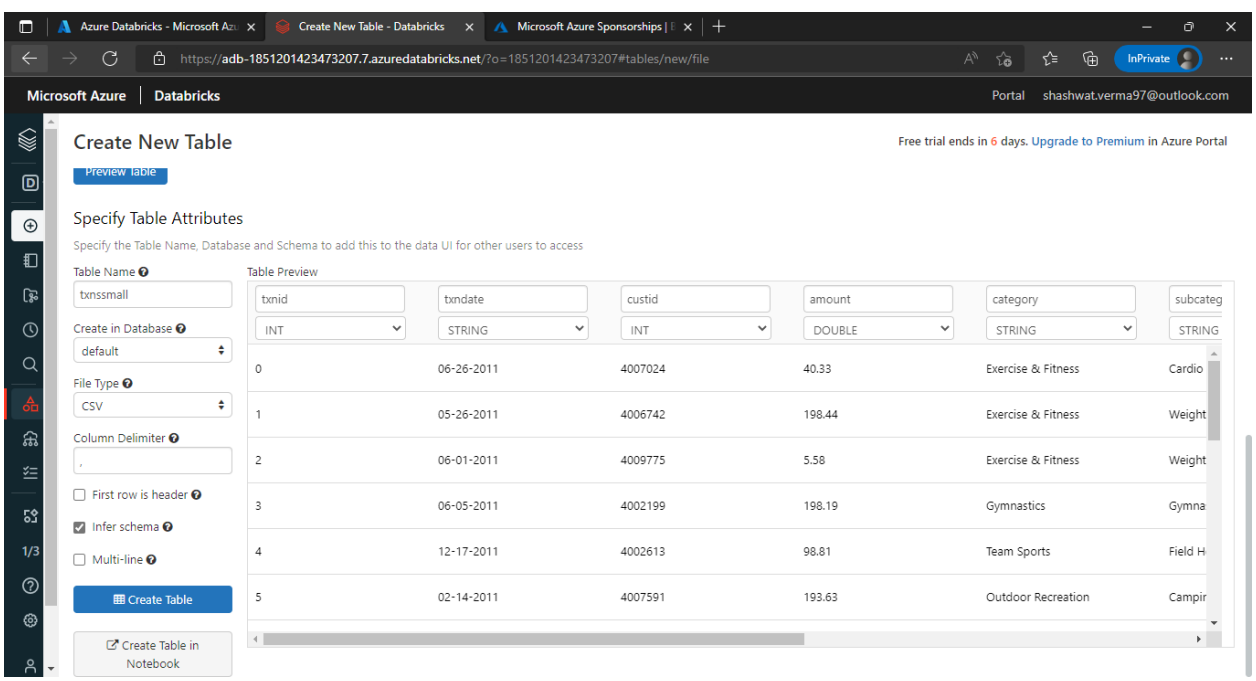
spark.sql("select category, SUM(amount) AS revenue FROM txnsdata GROUP BY category ORDER BY revenue LIMIT 1").show()



4. Table Method in Databricks

We created a Table using UI inside 'mytestcluster9706' and directly uploaded the file 'txnsSmall'

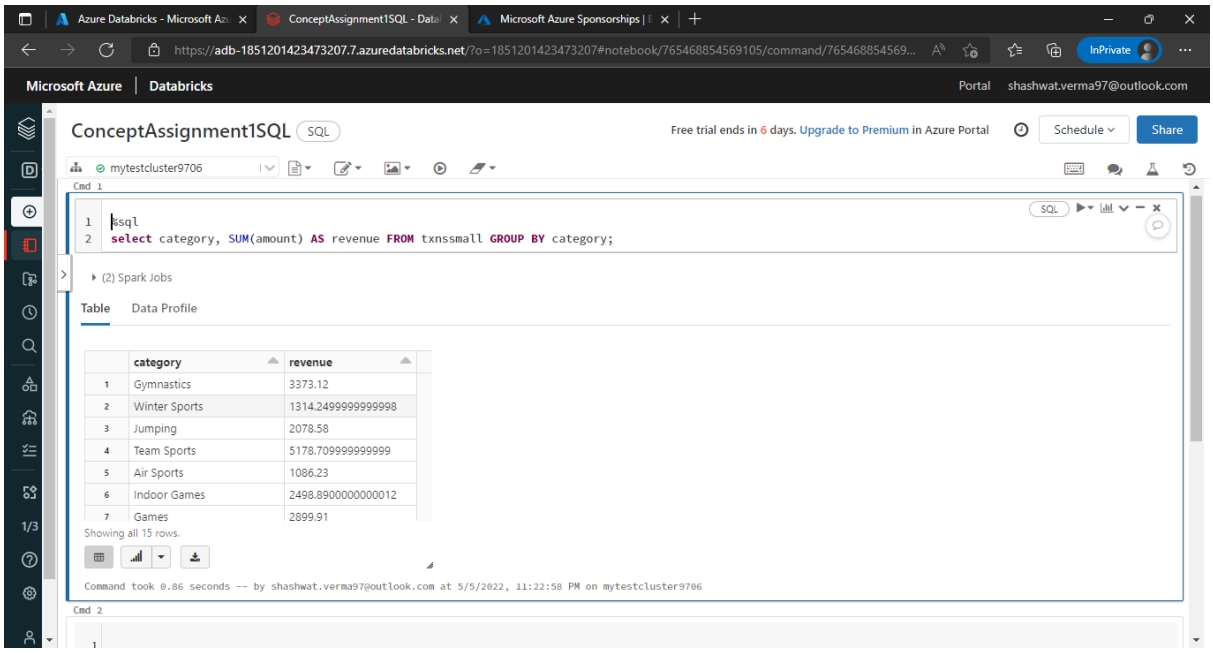




After hitting create table button for 'txnssmall' we created a new notebook 'ConceptAssignment1SQL' to perform the tasks

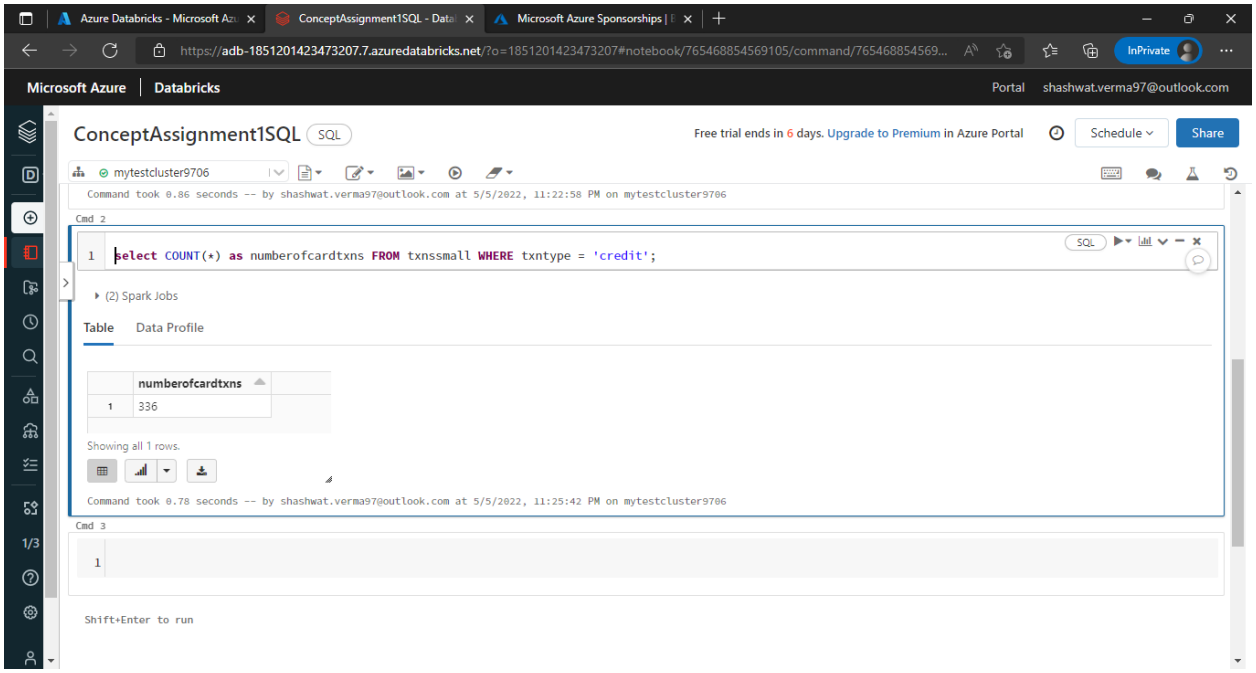
a. Find the total revenue generated based on category

select category, SUM(amount) AS revenue FROM txnssmall GROUP BY category;



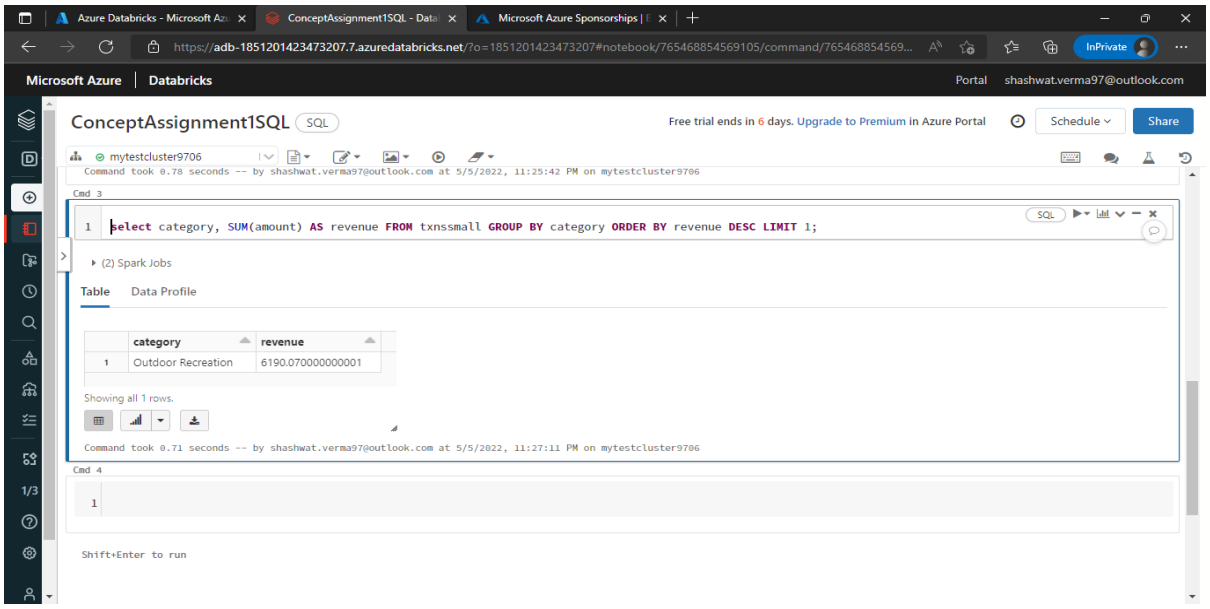
b. Find the total number of transactions done my card

select COUNT(*) as numberofcardtxns FROM txnssmall WHERE txntype = 'credit';



c. Find the highest selling category

select category, SUM(amount) AS revenue FROM txnssmall GROUP BY category ORDER BY revenue DESC LIMIT 1;



d. Find the lowest selling category

select category, SUM(amount) AS revenue FROM txnssmall GROUP BY category ORDER BY revenue LIMIT 1;