Graded Assignment – PySpark SQL

Business Scenario:

You have been provided with a credit card dataset containing information about credit card transactions. Your task is to analyze the dataset using PySpark SQL and derive meaningful insights. A couple of lists or reports are to be generated from this data as specified.

Data Sets and Data Dictionary:

The data is given in csv files. Below is the data dictionaries.

```
|-- RowNumber: integer (nullable = true)
|-- CustomerId: integer (nullable = true)
|-- Surname: string (nullable = true)
|-- CreditScore: integer (nullable = true)
|-- Geography: string (nullable = true)
|-- Gender: string (nullable = true)
|-- Age: integer (nullable = true)
|-- Tenure: integer (nullable = true)
|-- Balance: double (nullable = true)
|-- NumOfProducts: integer (nullable = true)
|-- IsActiveMember: integer (nullable = true)
|-- EstimatedSalary: double (nullable = true)
|-- Exited: integer (nullable = true)
```

Problem Statement:

Load the data of the files into PySpark SQL dataframes. Print the schema of the dataframe and show to first 10 rows of dataframe. ***- 10 Marks***

Now write queries to generate the following reports/lists. You can use PySpark SQL functions or queries with Spark Session object or both. You need to show the PySpark code & query and the top 10 rows from the results of the query in each case.

1. List the number of member who are eligible for credit card. ***- 10 Marks***
2. List the number of member who are eligible and active in the bank. ***- 10 Marks***
3. List the credit card users belonging to Spain ***- 10 Marks***
4. Save the above output as comma separated files. ***- 10 Marks***

You need to submit the HTML file or the IPYNB file with the details outlined above.