Spark SQL AirBnB

This assignment will use Spark core and Spark SQL to query a dataset. The dataset, originally sourced from Kaggle, contains information from airBnB about properties in five cities.

The data has the following columns:

id, name, host_id, host_name, neighbourhood_group, neighbourhood, room_type, price, minimum_nights, number_of_reviews,availability_365, city

Using Databricks, load the file AB_US_2020.csv from Moodle into a directory on Databricks:

/Filestore/tables/ass4

Q1. Spark Core

After loading the data, create a Notebook in Databricks (and start a cluster under Compute).

You can use the following code to load the data from the file into an RDD, ready to query:

```
import pyspark

sc = pyspark.SparkContext.getOrCreate()
inputRDD = sc.textFile("/FileStore/tables/ass4")
airbnbRDD = inputRDD.map(lambda x: ( x.split('|') ) )

# write your query here

display(resultRDD.collect())
```

Write the following queries. Place each question in a separate cell of the notebook.

a). How many properties are in each city

b). List number of properties in each neighbourhood in Los Angeles, in descending order.

c). List the number of property types (room_type) in each neighbourhood in Seattle.

d). List the average price of an 'Entire home/apt' in each City.

Q2. Spark SQL

This section will use Spark SQL to query the dataset. Use the following code to load the data into a Dataframe.

```
import pyspark
import pyspark.sql.functions as sqlFunc

# Schema Definition
my_schema = pyspark.sql.types.StructType([pyspark.sql.types.StructField("id",
pyspark.sql.types.IntegerType(), False),
                                          pyspark.sql.types.StructField("name",
pyspark.sql.types.StringType(), False),
                                          pyspark.sql.types.StructField("host_id",
pyspark.sql.types.IntegerType(), False),

pyspark.sql.types.StructField("host_name", pyspark.sql.types.StringType(), False),

pyspark.sql.types.StructField("neighbourhood_group",
pyspark.sql.types.StringType(), False),

pyspark.sql.types.StructField("neighbourhood", pyspark.sql.types.StringType(),
False),

pyspark.sql.types.StructField("room_type", pyspark.sql.types.StringType(), False),
                                          pyspark.sql.types.StructField("price",
pyspark.sql.types.IntegerType(), False),

pyspark.sql.types.StructField("minimum_nights", pyspark.sql.types.IntegerType(),
False),

pyspark.sql.types.StructField("number_of_reviews", pyspark.sql.types.IntegerType(),
False),

pyspark.sql.types.StructField("availability_365", pyspark.sql.types.IntegerType(),
False),
                                          pyspark.sql.types.StructField("city",
pyspark.sql.types.StringType(), False)
                                         ]
                                       )

# Load data from file using schema
inputDF = spark.read.format("csv") \
                    .option("delimiter", "|") \
                    .option("quote", "") \
                    .option("header", "false") \
                    .schema(my_schema) \
                    .load('/FileStore/tables/ass4')
```

Create the following queries using the inputDF dataframe loaded above:

1). Select the property type and city for all rooms in Hawaii

2). Select number of long term rentals (with minimum_nights > 180) in each city.

3). Maximum, minimum and average price of each room_type in Rhode Island.

When you have completed the queries above in a Notebook, select the file icon at the top of the Notebook and select Export -> iPython Notebook.

Upload your notebook to Moodle.