Assignment 9

You are practicing Spark DataFrames in the practice project

https://www.scriptedin.com/contests/view/40

(10 points) In this assignment, you will get familiar with programming using Spark DataFrames. You are provided with a starter notebook that read data for you into 8 DataFrames.

1. (1 point) Create a DataFrame based on the payment DataFrame. It contains a list of customer_id and total payment amount grouped by customer_id, sorted by the total amount in descending order. Other columns may be included as needed. Top 3 should be as follows for self-verification purpose

```
+-----------+------------------+
|customer_id|       sum(amount)|
+-----------+------------------+
|        526| 221.5500000000001|
|        148| 216.5400000000001|
|        144|195.58000000000007|
+-----------+------------------+
```

2. (1 point) Create a DataFrame based on the rental DataFrame. It contains a list of rentals from the rental DataFrame that were rented in 2005 ordered on return_date in ascending order. Other columns may be included as needed. Top 3 should be as follows for self-verification purpose

```
+---------+-------------------+-----------+-----------+-------------------+--------+-------------------+
|rental_id|        rental_date|inventory_id|customer_id|        return_date|staff_id|        last_update|
+---------+-------------------+-----------+-----------+-------------------+--------+-------------------+
|       32|2005-05-25 04:06:21|        3832|        230|2005-05-25 23:55:21|       1|2006-02-15 21:30:53|
|       21|2005-05-25 01:59:46|         146|        388|2005-05-26 01:01:46|       2|2006-02-15 21:30:53|
|       14|2005-05-25 00:31:15|        2701|        446|2005-05-26 02:56:15|       1|2006-02-15 21:30:53|
+---------+-------------------+-----------+-----------+-------------------+--------+-------------------+
```

3. (1 point) Create a DataFrame based on rental DataFrame . It contains a list of rental_id count from rental DataFrame grouped by staff_id. Result should be as follows for self-verification purpose.

staff_id	count
1	8040
2	8004

4. (1 point) Now add staff first and last name. Result should be as follows for self-verification purpose.

first_name	last_name	count
Mike	Hillyer	8040
Jon	Stephens	8004

5. (2 points) Create a DataFrame based on rental DataFrame, joined with inventory, store, and address DataFrames to get store address, joined with customer DataFrame to get customer first and last name, joined with staff DataFrame to get staff first and last name, joined with film data set to get film name and description, and ordered by return_date in the ascending order. The 3 first row should be as follows for self-verification purpose

rental_date	return_date	address	title	description	first_name	last_name	first_name	last_name
2005-05-24 22:53:30	2005-05-26 22:04:30	47 MySakila Drive	BLANKET BEVERLY	A Emotional Docum...	CHARLOTTE	HUNTER	Mike	Hillyer
2005-05-24 22:54:33	2005-05-28 19:40:33	28 MySQL Boulevard	FREAKY POCUS	A Fast-Paced Docu...	TOMMY	COLLAZO	Mike	Hillyer
2005-05-24 23:03:39	2005-06-01 22:12:39	28 MySQL Boulevard	GRADUATE LORD	A Lacklusture Epi...	MANUEL	MURRELL	Mike	Hillyer

6. (1 point) Create a DataFrame that contains a list of customer_id whose sum amount of payment in 2005 is more than $200. Result should be as follows for self-verification purpose.

customer_id	sum(amount)
526	221.5500000000001
148	216.5400000000001

7. (1 point) Join the result with customer DataFrame set to get their names. Result should be as follows for self-verification purpose.

```
+----------+---------+-----------------+
|first_name|last_name|      sum(amount)|
+----------+---------+-----------------+
|      KARL|     SEAL|221.5500000000001|
|   ELEANOR|     HUNT|216.5400000000001|
+----------+---------+-----------------+
```

8. (2 points) Create a DataFrame that contains a list of customer_id that rented in 2005; Create another DataFrame that contains a list of customer_id that rented in 2006 (filter using rental_date).
   a. (1 point) Now create a DataFrame of customer_id that exist in the 2005 DataFrame but not in the 2006 DataFrame . The count should be 441.
   b. (1 point) Now create a DataFrame of customer_id that exist in the 2006 DataFrame but not in the 2005 DataFrame . The count should be 0.

To get full credit, you are required to run these on (1) **a local computer** (2) **Google Colab and Google Drive** (3) **Google Cloud Dataproc and Storage using the free credit provided in the previous lecture (turn off all your resources after the assignment)**. Include snapshots of the screen to prove it. Missing each minus 1 point

Note: Use the provided template notebook. It solves an issue with Java. Google now has Java JDK 11 in Colab, which is not compatible with Spark (Spark can use up to Java 8). This notebook tries to install Java 8 and Spark 2.2.3.

The amount column may need casted (see the csv file in Notepad), for example:

payment_df.withColumn("amount", payment_df["amount"].cast("double"))

Submission on Husky includes a zip file of:

a. A notebook named as “group_xxx_assignment_yyy.ipynb”

b. A Word document/article with detailed explanation

c. All the other files

Also archive the notebook and the article (only after the deadline) to the project site.