Assignment 8

You are practicing RDD transformations and actions in the practice project

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

Use the attached payment.csv in the assignment. Write code to create RDDs for the following. For each RDD, print a few records:

1. (2.5 points) Average amount of payment by month & year. Example:
   ```
   ('200505', 4.169775280898756)
   ('200506', 4.166038062283568)
   ('200507', 4.22796751601904)
   ('200508', 4.232834534904548)
   ('200602', 2.8251648351648404)
   ```
2. (2.5 points) Total payment by customer id. Example:
   ```
   ('1', 118.67999999999992),
    ('2', 128.72999999999993),
    ('3', 135.73999999999998)
   ```
3. (2.5 points) Total payment by staff id. Example:
   ```
   ('1', 33489.470000005174),
   ('2', 33927.04000000482)
   ```
4. (2.5 points) Average payment by month & year and by customer id. Example:
   ```
   (('200505', '1'), 1.9900000000000002),
    (('200506', '1'), 4.561428571428571),
    (('200507', '1'), 4.240000000000001),
    (('200508', '1'), 2.899090909090909),
    (('200505', '2'), 4.99),
    (('200506', '2'), 2.99),
    (('200507', '2'), 5.418571428571428),
    (('200508', '2'), 4.080909090909092)
   ```

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 notebook of the 2 applications (Morning Star Ratings and Average Price by Month) in HuskyCT to start with as 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.

Also open the csv file using Notepad or similar, not Excel as date columns may look different. Also note the double quotes are in the data and need to be processed.

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.