ACADGILD

Session 46:
RDD & Dataframes
Assignment 1

Assignment 1 – You must perform the given tasks.

Table of Contents

1.	Introduction	3
2.	Objective	3
3.	Prerequisites:	3
4.	Associated Data Files	3
5.	Problem Statement	3
6.	Expected Output	3
7.	Approximate Time to Complete Task	3

1. Introduction

In this assignment, you need to perform the given tasks.

2. Objective

This assignment will help you to consolidate the concepts learnt in the session 4.

3. Prerequisites:

None

4. Associated Data Files

None

5. Problem Statement

Task 1

Given a dataset of college students as a text file (name, subject, grade, marks) :
Dataset
Problem Statement 1:
1. Read the text file, and create a tupled rdd.
2. Find the count of total number of rows present.
3. What is the distinct number of subjects present in the entire school
4. What is the count of the number of students in the school, whose name is Mathew and marks is 55
Problem Statement 2:
1. What is the count of students per grade in the school?
2. Find the average of each student (Note - Mathew is grade-1, is different from Mathew in some other grade!)
3. What is the average score of students in each subject across all grades?
4. What is the average score of students in each subject per grade?
5. For all students in grade-2, how many have average score greater than 50?
Problem Statement 3:
Are there any students in the college that satisfy the below criteria :
1. Average score per student_name across all grades is same as average score per student_name per grade
Hint - Use Intersection Property.

Task 2

1) What is the distribution of the total number of air-travelers per year
2) What is the total air distance covered by each user per year
3) Which user has travelled the largest distance till date
4) What is the most preferred destination for all users.
Use below link to download the dataset:

https://drive.google.com/drive/folders/0B_P3pWagdIrrVThBaUdVSUtzbms

Task 3

1) Which route is generating the most revenue per year
2) What is the total amount spent by every user on air-travel per year
3) Considering age groups of < 20 , 20-35, 35 > ,Which age group is travelling the most every year.

Use the dataset given below:
https://drive.google.com/drive/folders/0B_P3pWagdIrrVThBaUdVSUtzbms

6. Expected Output

Solution report with commands, explanation to commands and screenshot for output.

Report shall be in PDF format. Submitted in GitHub

7. Approximate Time to Complete Task

8 hours.