**Assignment -44**

**Task 1-Problem Statement-1**

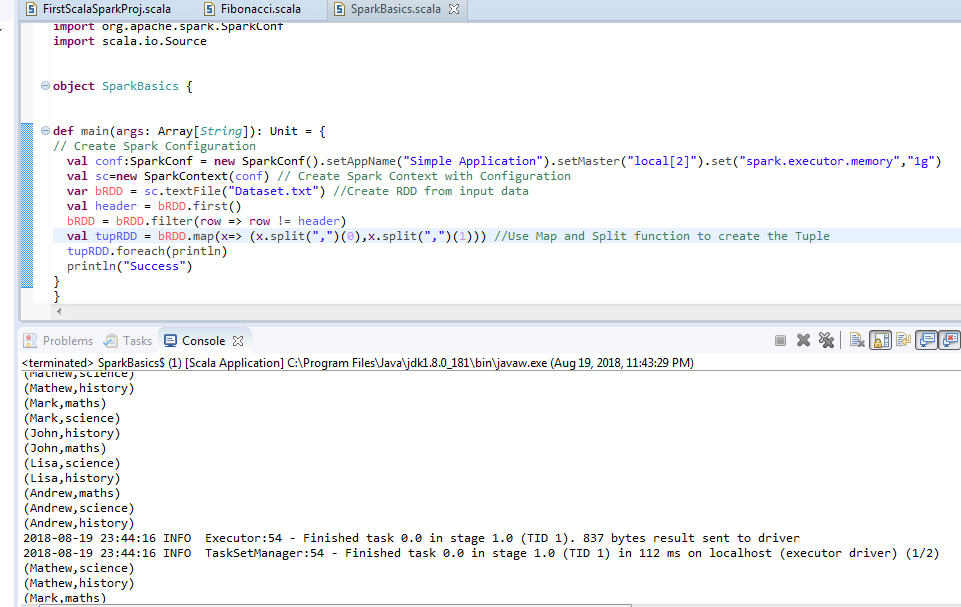
**ALL TASK-1 Code available in the file SparkBasics.scala in GitHub**

Given a dataset of college students as a text file (name, subject, grade, marks) :

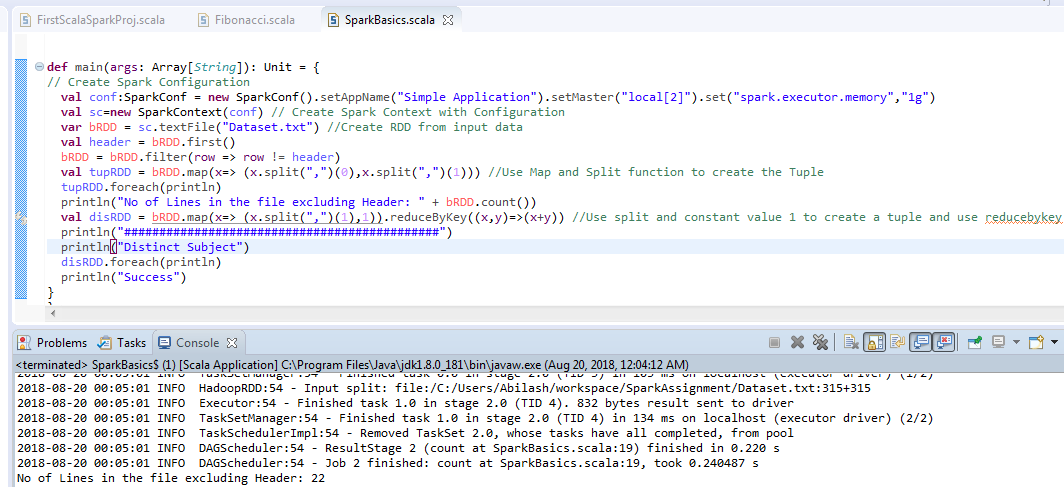
Dataset

Problem Statement 1:

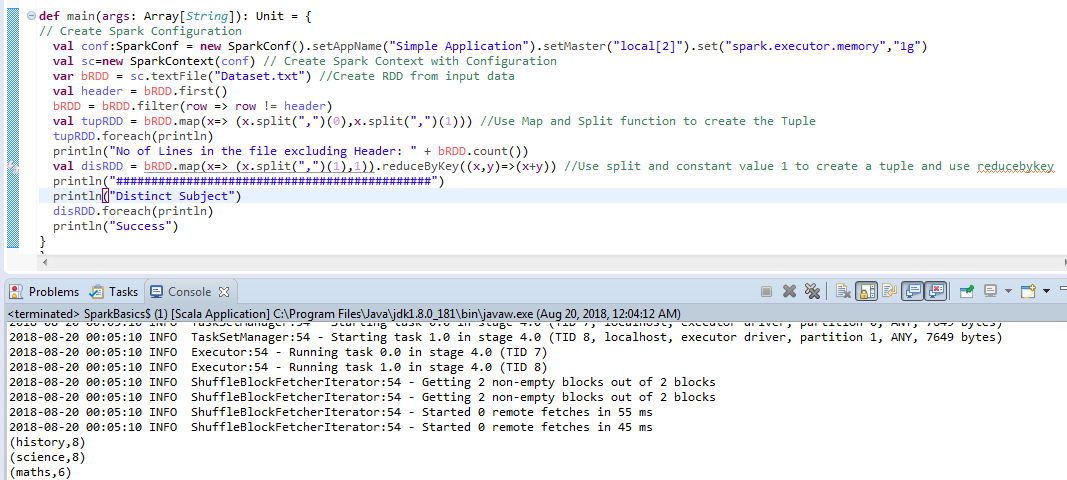
1. Read the text file, and create a tupled rdd.



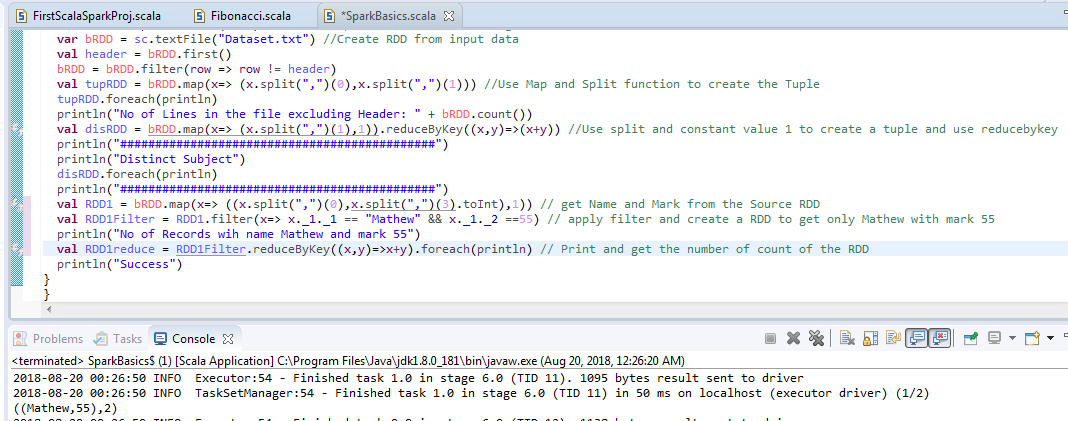
1. Find the count of total number of rows present.



1. What is the distinct number of subjects present in the entire school

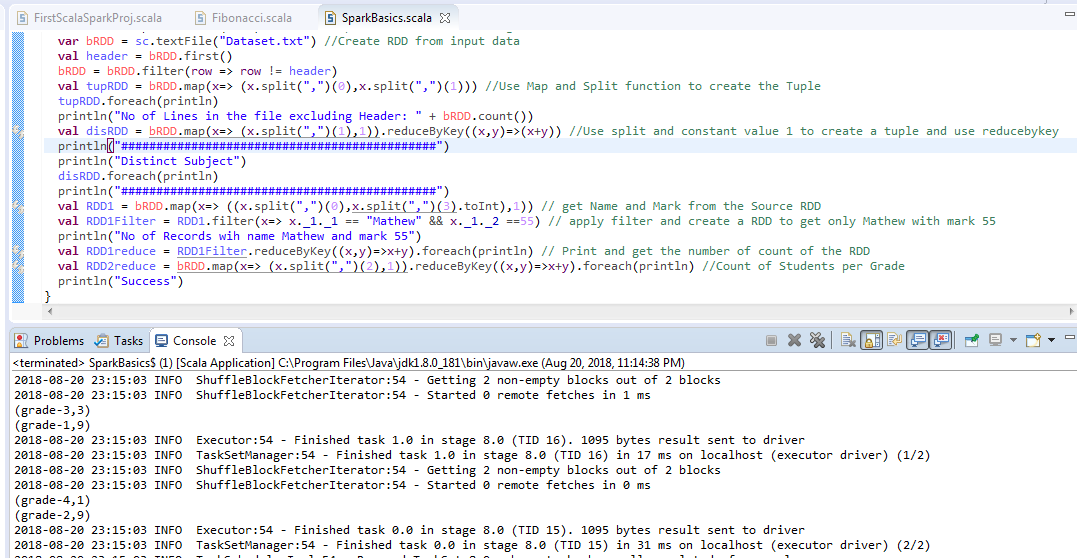


1. What is the count of the number of students in the school, whose name is Mathew and marks is 55?



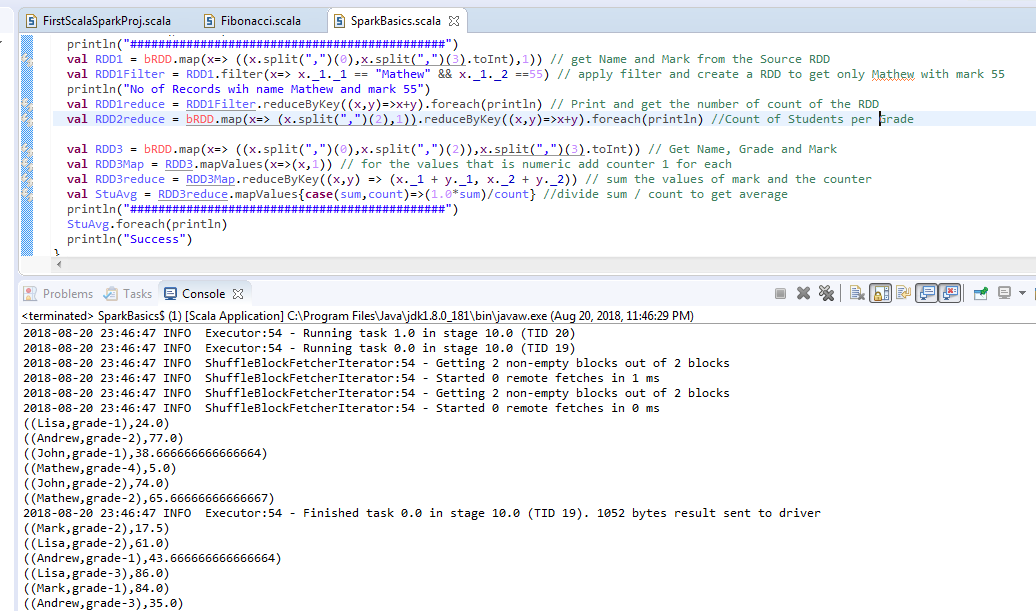
**Task-1 Problem Statement-2**

1. What is the count of students per grade in the school?

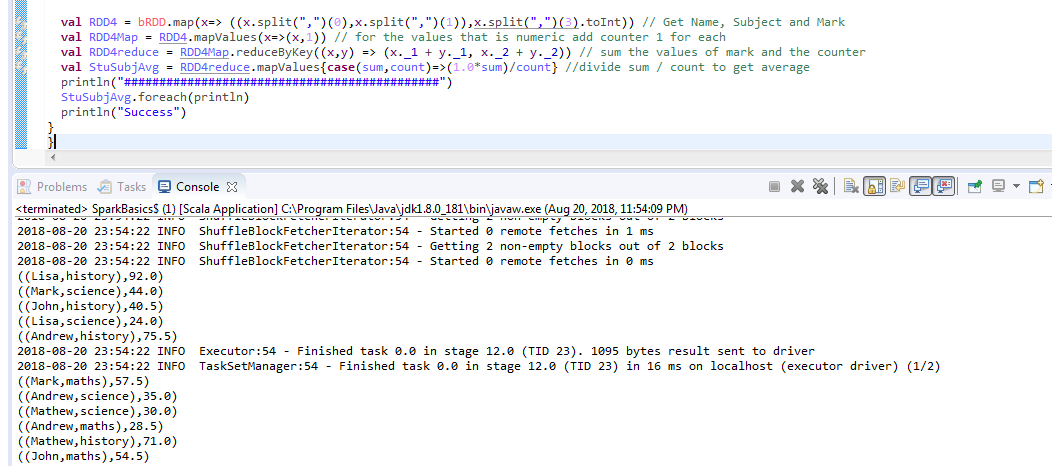


1. Find the average of each student (Note - Mathew is grade-1, is different from Mathew in

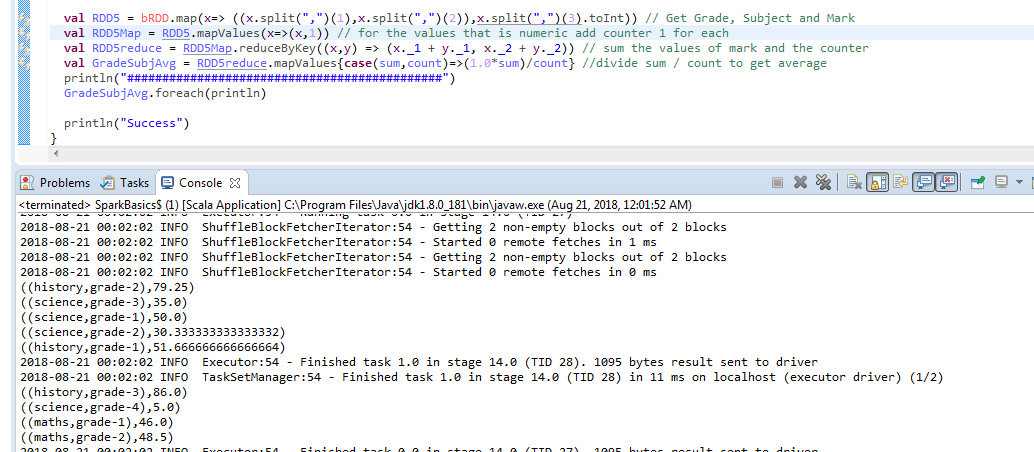
some other grade!)



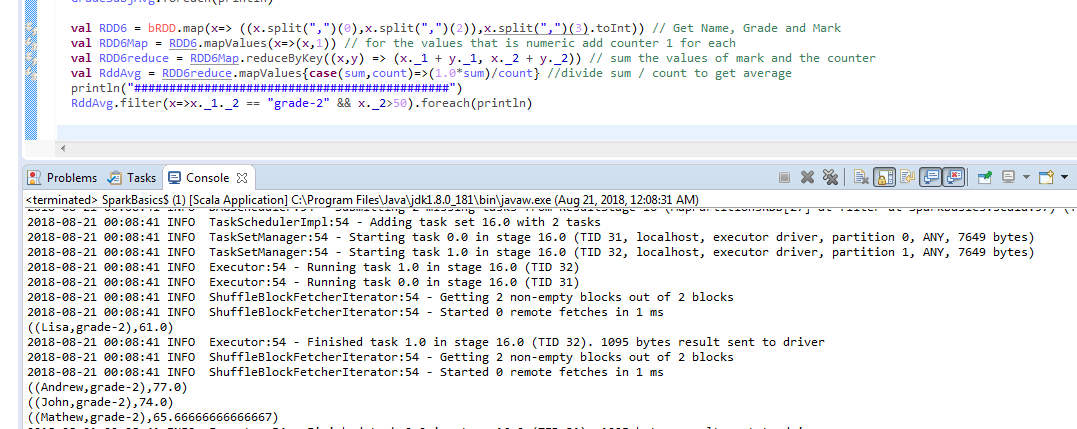
1. What is the average score of students in each subject across all grades?



1. What is the average score of students in each subject per grade?



1. For all students in grade-2, how many have average score greater than 50?

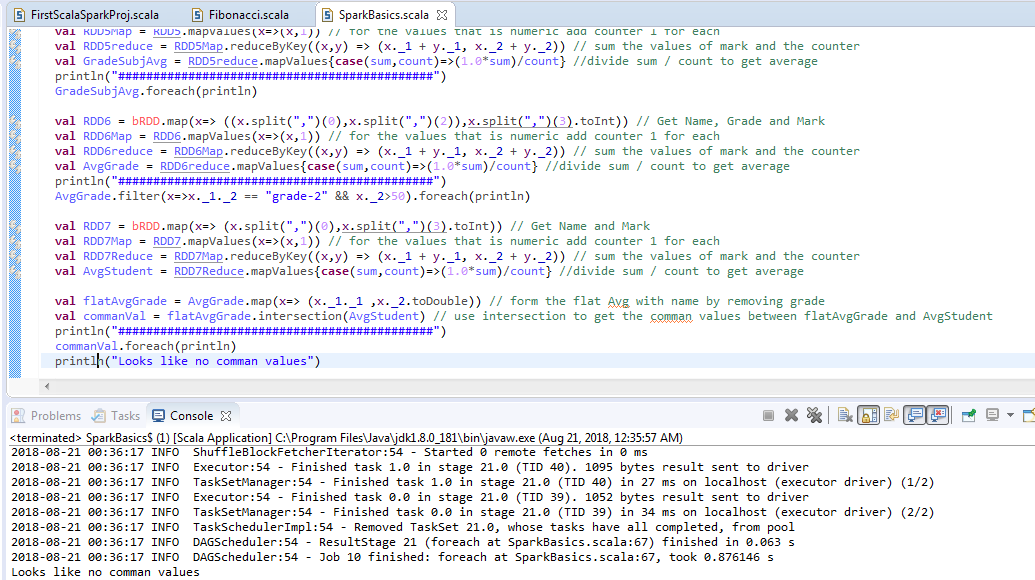


**Task-1 Problem Statement 3**

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.

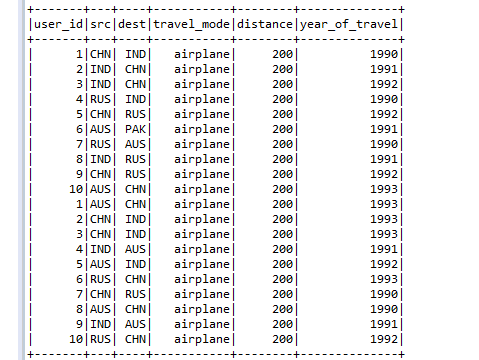


**Task-2**

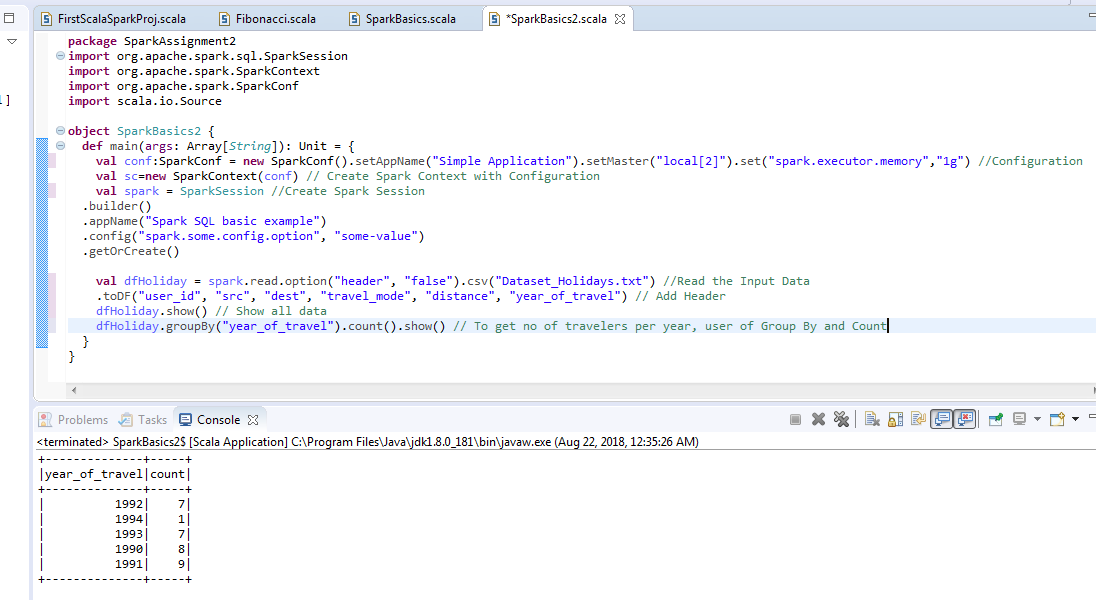
**ALL TASK-1 Code available in the file SparkBasics2.scala in GitHub**

1. What is the distribution of the total number of air-travelers per year

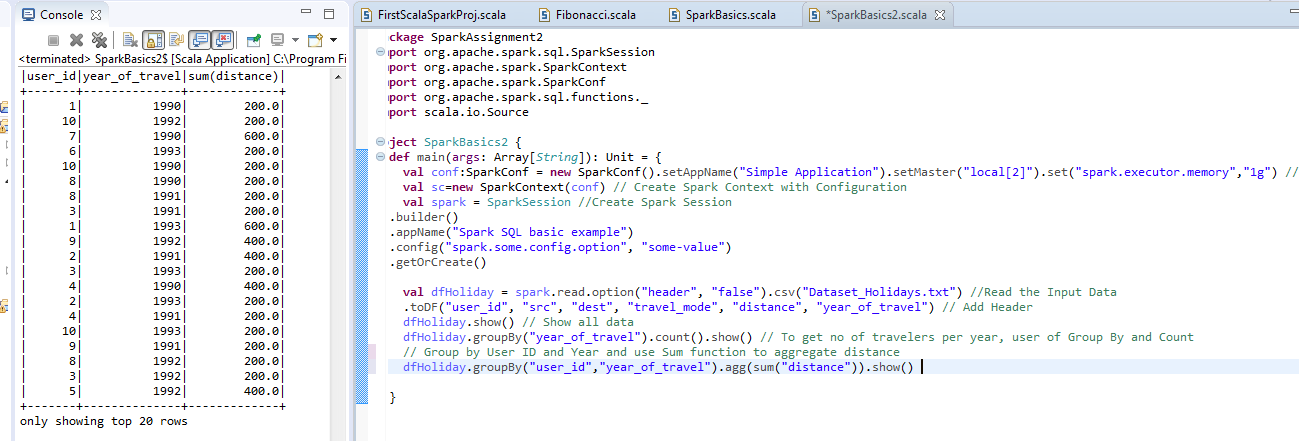
**ALL DATA**



**Code and the Output for the question 1**



1. What is the total air distance covered by each user per year



1. Which user has travelled the largest distance till date?



**Here two users have travelled more distance.**

1. What is the most preferred destination for all users?



**Here IND is the most preferred destination as it where more people has visited.**