PySpark RDD Assignment

Given two separate datasets of a sports complex with the following shemas:

Cust ID	First Name	Last Name	Age	Profession

Trans ID	Date	Cust ID	Cost	Game	Equipment	City	State	Mode

Load data from **trans240.csv** and **cust.csv** and perform the following queries:

1. For each month, show the number of distinct players, and the total cost, the results are sorted by moth
2. For each month, show the name three youngest player, the results are sorted by the month
3. For each state, show the number of distinct players of each state, the results are sorted by the number of players
4. For each state, show the name of three oldest player in each state, the results are sorted by the state
5. For each state, show the average age of players in each state, the results are sorted by the average age
6. For each player ID, show the average number of game per month, the results are sorted by player ID
7. For each player ID, show the game with highest total cost, and the total cost of this game, the results are sorted by player ID
8. For each player ID, show the month with highest total cost, and the total cost in this month
9. For each player ID, show the list of three games with most transactions, and the list of the number of transactions of these three games, the results are sorted by player IDs
10. For each game, show the number of transactions, the results are sorted by game
11. For each game, show the number of transactions, and the total cost, the results are sorted by game
12. For each month, show the number of transactions, and the total cost, the results are sorted by month

Notes:

1. Provide the information of your team including team number and list of member.
2. Submit only after finishing all queries. You have 7 days to submit your works on courses.uit.edu.vn
3. Each team has one member submiting three files:
   a. .IPYNB notebook file with all results (for checking if necessary)
   b. .PDF file printed from notebook with all results (for quick check)
   c. .TXT file containg codes of all queries (for Plagiarism check)
4. Do not copy each other. Teams with similar codes (checked by software) will get a zero point in this assignment.