

Introduction to Data Science

Final Project

The main goal of this project is to practice and apply what you have learned to real-world tasks.

1. No cheating. If any it will be hardly penalized for both parties.

2. Student must choose their groups; each group should not consist of more than four students (1-4) Please submit group members using the next link:

(https://docs.google.com/forms/d/e/1FAIpQLScqN7PnX3eQ1UfpoIGROWkapvNIvgfDIogdtqxTCY3MGRMYNQ/viewform?usp=sf_link).

- Deadline for submitting this form is on 16-12-2021 at 11:59 p.m.
- After the registration form closes, each group will be assigned to a certain number that will be announced on teams.
- Discussion of the projects starts on 1-1-2022 (According to the timetable that will be announced later).

3. Each group must prepare a **pdf report** call “**Project Report + Group No.**”, the report must contain at least the next items:

- Student’s name, ID, and Group.
- Explain the problem and briefly describe the role of each member. **Note that:** the problem description must answer the following questions:
 - a. What will the program do?
 - b. What the input to the program will be.
 - c. What the output from the program will be.
- The full description of your dataset.

- Screenshots from your Project steps.
- Explain your results and insight by describing your plotted graphs.
- Discussing every line in the Code (libraries used + attributes)
✓ (Screenshot for code parts + Describing what it does)

4. We have Grocery (GRC) dataset where you can download from [Click here to download the dataset.](#)

Using this dataset, you are asked to Use (R) to do the following tasks:

- a. Assess and clean your data if needed
- b. Use a different type of Data Visualization tools for each of the following:
 - i. Compare cash and credit totals.
 - ii. Compare each age and sum of total spending.
 - iii. Show each city total spending and arrange it by total descending.
 - iv. Display the distribution of total spending.
- c. Put all previous plots in one dashboard.
- d. Split the customers to (n) groups using one of the studied methods (n will be user input) according to the sum of total spending and their ages and print a table displaying each customer name, age, total and the computed cluster number.
- e. Generate association rules between items with minimum support and confidence taken from the user inputs (State the algorithm used).

5. You can use the following guidelines to assist you in implementing your Program:

❖ **Program user inputs**

Variable Name	Label	Notes	Validation
datasetPath	Dataset path	User should input the full path of the csv file	Required
numberOfClusters	Numbers of clusters	To use in the k-means function	Number between 2 and 4
minSupport	Minimum Apriori support	To use in the Apriori algorithm	Number between 0.001 and 1
minConfidence	Minimum Apriori confidence	To use in the Apriori algorithm	Number between 0.001 and 1

Submission Details:

You should submit the following on teams according to instructions:

- ❖ R Script of the code.
- ❖ Project Report (File name must be Project_report + group number EX: "Project_Report_17.pdf")

Good Luck😊