**HIGH-DIMENTIONAL ANALYSIS**

**---**

**Programming Exercise – 03**

Deadline: 23h59 11/06/2022

Submitting via email: [dxtien95@gmail.com](mailto:dxtien95@gmail.com)

Programming language: Python

**PROBLEM:**

Data: <https://drive.google.com/file/d/1IqSv-q8bE3Fa5n93ZB7-Jza0DffC3TK4/view?usp=sharing>

Import the CSV file *'College.csv'* and do the following analytics:

1. a) Find all private schools with a graduation rate of higher than 90%.
2. b) Using all attributes from the initial dataset except for the Private attribute, do the K-mean clustering method with two clusters:
   * + - 1. 1. Implement k-means clustering from scratch
         2. 2. Apply sklearn’s k-means
3. c) Compare the results of your k-means and sklearn with the Private attribute (ground-truth).

**Dataset description:**

* Statistics for a large number of US Colleges from the 1995 issue of US News and World Report, with 777 observations on the following 18 variables.

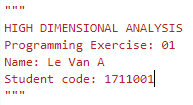
1. Private**:** A factor with levels No and Yes indicating private or public university
2. Apps: Number of applications received
3. Accept: Number of applications accepted
4. Enroll: Number of new students enrolled
5. Top10perc: Pct. new students from top 10% of H.S. class
6. Top25perc: Pct. new students from top 25% of H.S. class
7. F.Undergrad: Number of fulltime undergraduates
8. P.Undergrad: Number of parttime undergraduates
9. Outstate: Out-of-state tuition
10. Room.Board: Room and board costs
11. Books: Estimated book costs
12. Personal: Estimated personal spending
13. PhD: Pct. of faculty with Ph.D.’s
14. Terminal: Pct. of faculty with terminal degree
15. S.F.Ratio: Student/faculty ratio
16. perc.alumni: Pct. alumni who donate

**SOURCE**

* This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University.The dataset was used in the ASA Statistical Graphics Section’s 1995 Data Analysis Exposition.

After finishing your task, please write a short report or a summary (pdf file) to explain your answers, ideas and the way your code works.

**NOTICE**

1. Please send the two files (coding and report file) before the due date. Or send the jupyter-notebook file (ipynb, html, pdf) or google-colab link.
2. The mail subject and the folder’s name would be *[HDA2020\_PEXX\_Name\_StudentCode]*, where PE means Programming Exercise. For example: *HDA2020\_PE01\_Le\_Van\_A\_1711001* or *HDA2020\_PE01\_LeVanA\_1711001*.
3. Inside the coding file, there should be a brief introduction (as example below). 
4. There is **NO** acceptance for **cheating** or **copying**.

**TUTORIAL**

Export html file from jupyter-notebook

