Homeworks in unsupervised learning.

Please take the census data:

https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29

and please, do the following:

1) Remove the columns dAge, dHispanic, iYearwrk and iSex from the data, and treat those as external variables.
2) Cluster the remaining columns, and explain what are is the best clustering method.
3) Associate the clusters with the external variable, and explain what external variable is best associated with clusters, and what clustering method best associates with the outside variables.
4) Find anomalies in the data, and test whether anomalies are associated with any of the external variables.
5) Reduce the dimension of the data and propose a visualization that best characterize the clusters associated with the external variables. Please find a smart presentation scheme to highlight both clusters and variables.

You have until 4/3/2022 at 12:00 to submit the HW.

The grade will be 60 % on the quality of the analysis and the statistical tests and 40% on the presentation.

Good luck

Instructions for submissions. Please follow those carefully. Your note depends on it.

A) The submission must be in tex (I suggest overleaf it is easy to use)
B) The structure should be title, abstract, introduction, methods, results, discussion. References
C) Please use bibtex for references
D) Please have a git with your entire code and results, and a link to the git at the end of the abstract
E) The total length is limited to 8 pages.
F) The intro must contain a description of what you plan to do and why you do it
G) The methods should contain all the technical details of what you do (methods, algorithms, parameters, train test split, statistical tests….). **No need** to explain what the algorithms do, only what is specific to the application you did.
H) There are no figures in the intro and methods
I) The results are all your results. Each figure and table require a detailed enough figure legend that the table/legend can be understood directly.
J) Every figure/table should be there to answer a claim. For example, Algorithm 1 is better than 2 (Figure XXXX) T test p<XXXXXXX).
K) Please try to have as informative figures and tables as you can.
L) All figures should be high quality – all axes should have captions. Please do not put titles to plot sub/plots (and just for OCD like me, please start sentences and captions with Capitals).
M) Try to have multiple subplot in each figure, so that when one sees a figure, he can understand everything you have to say about a claim.
N) Use a font size of 11 or 12 (better to use times new roman).
O) The discussion should be at least 0.5 pages and have some smart insight about the results.