DM Meeting Questions

1. hello! If we filter the rows by the Recency of Giving (containing lapsed donors) on the MDMAUD column, it reduces the rows in the df to less than 300. My question is: am I thinking in a wrong way to start do the clusters or is this a very little effort giving the df inicial size
2. If we are doing the clustering with PCA is there a way to go back to our variables so we can interpret them easily?
3. Pandas Profiling will tell you exactly which are metric or categorical
4. What should we consider a good/acceptable R^2 for our cluster solutions in this particular dataset?
5. The code for the outliers in lab 7 is not working for our dataset. Clusters analysis can replace that method ?
6. So we have a really raw data ! Sometimes with lack of information also… In machine learning we talked that dropping like 3/4% of the data was acceptable but should we drop more in this case ? We have a lot of donors !! Or reaching the "worst" 4% we should start to make some assumptions and try to treat some weird/missing values?
7. <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html>
8. Hello professors! In the outlier removal with the IQR is it normal to have 6 as the multiplier? Because if we define the default 1.5 most of the observations are considered outliers.
9. On the forum it was the contrary approach, to keep all as lapsed donors
10. Hi, if we are not using categorical variables in the clustering itself and only for interpretation, why should we one hot encodinng them?
11. The purpose of running K-Means before applying hierarchical clustering is just to define the initial cluster number for the hierarchical right?
12. Can you clarify a bit the desired report structure?
13. Age seems a very relevant feature, even if DOB misses a considerable % of data. Is it reasonable to fill the missing data with some mean value, global, or possibly grouped by which war their veterans of?
14. Is it ok to consider one time doners as an own cluster or how to we deal with features having missing values by definition for those doners (eg 'time\_between\_donations')?
15. About the bottom-up approach: We start with a low number of variables (like 15-20), but until what point should we keep adding more variables? Is there like a maximum number of metric features to include in our clustering?
16. So we should keep columns that have a lot of missing values? There are 66 initial columns that have >40% missing values. That's about 14% of the raw dataset
17. do we need to put the code in GitHub and give you a link for it?
18. is it posible to apply the algorithm only to a random sample of the data? say 25%
19. When we try KNNImputer to impute the missing values, should be consider the features that is correlated to the imputed one or should we include every feature? Can we just pick the feature subjectively, e.g. impute INCOME we only use WEATH and AGE and amount of donation?
20. You mentioned to have a data pre-processing section in the report, in this part would you like us to justify the reason why we drop each of the features we dropped or shall we put that in the appendix instead (since we dropped a loooot of problematic variables)?
21. when caracterizing the clusters, what does happen if we don't find big differences between clusters?
22. Following Nadine question, for the Neighborhood features, there are like 283 features, so initially we just pick the best one out subjectively and start clustering the chosen only, is it acceptable?
23. could we use PCA for feature selection? since, there are some variables that won't be important for PCA
24. Can we drop the variables that are clearly useless? like the time to work, which transportation method (and sure this may give socio economic info, but for that there are better ones like average income), because we have been told not to disregard variables by subjective means, but it is insane to keep all of those, being that some don't drop by objective methods
25. what is the threshold for highly correlated features to drop? 80% correlation?
26. because I am really getting the impression that the professors have different opinions on how to approach this. Joao seems to be telling us to keep only variables that make sense and that we can explain how we are keeping those (which is much more in line with what we are told in theoretical classes). David seems to be telling to only drop the ones we can explain objectively why we drop them.
27. sorry, probably I wasn't clear because you have just answered like we can drop 66 columns with a lot of NaN, but my point is - they are not NaN's - e.g. person who didn't send a gift has nothing in RDATE and RAMNT - but it doesn't mean we have missing data here - it's just mean he sent nothing, no?
28. we had a meeting with David and we are following the top down approach analyzing all features and not assuming any judge on it in order to let the data speak for itself we are using mahalanobis and it’s giving 20% of outliers we understand that it was reasonable however how can we approach to not have more than 3-4% outliers?
29. low correlation means 20%?? 40%? Joao you mentioned to use the bottom-up approach for the features our clusters but David mentioned to use different sets of features to have different cluster topics/characteristics…I’m confused what to do?