Steps for running the NMF

Step 1: Reading in the data, note that hourly_radon is somewhat raw since it contains a lot other unnecessary variables like those related to oil&gas production (which contains more NA than other variables), so I filtered out those variables first then did na.omit

Step 2: split the data into VOCS (16 variables) and non-VOCS (6 variables). Both dataset should contain 2703 rows with no-na.

Step 3a: Preprocess the VOCs. First perform adjusting_neligible_background_from_LOD, this requires a vector of all background levels of the variables and their LODs. We further splited the VOCs based on their LOD values (since ethane, propane, benzene, and acetylene have different LODs)

Step 3b: After adjusting for background levels, we did replace_negative_with_random. Again this has to be done separately based on their LODs.

Step 3c: Merging all VOCs together.

Step 4: Normalize the non-vocs using normalize_column.

- the file NMF_voc_norm.Rmd also normalized the VOCs whereas NMF.Rmd simply uses the un-normalized VOCs.

Step 5: Combine the normalized non-vocs and VOCs together, obtain their transpose. Created a LOD_vector which is used for creating the weight / uncertainty matrix. The uncertainty matrix is created based on the Guha paper.

Step 6: set a seed and ran the estimate rank with a range of rank to test from 4 to 20, method of , and number of run = 5.

Step 7: Based on the above result, rerun the nmf but with the optimal rank and plot their source contributions.

Note: it seems that NMF with normalized-VOCs produces a much more meaningful source contribution plots.